DOCUMENT RESUME

ED 181 049 TM 009 951

TITLE	Proceedings of the Invitational Conference on Testing Problems (1st, New York, New York, October 30, 1948). "Validity, Norms and the Verbal Factor".
INSTITUTION	Educational Testing Service, Princeton, N.J.
PUB DATE	30 Oct 48
NOTE	113p.
EDRS PRICE	MF01/PC05 Plus Postage.
DESCRIPTORS	Academic Achievement: Achievement Tests: Aptitude Tests: Behavioral Objectives: Community Characteristics: Course Content: Elementary Grades: Elementary Secondary Education: Individual Differences: Intermediate Grades: *Norms: Private Schools: Problem Solving: Public Schools: Scores: Secondary School Students: *Test Construction: *Testing Problems: Test Items: *Test Validity: *Verbal Ability: Vocabulary Skills
IDENTIFIERS	Air University AL

ABSTRACT

The conference is divided into three sections: (1) constructing examinations so that they will be valid measures of important functions; (2) defining the nature of the verbal factor as used in aptitude test batteries; and (3) establishing the type of norms most useful and important for the interpretation of achievement test scores. Papers are entitled: Item Selection to Maximize Test Validity: Criticisms of Commonly Used Methods of Validating Achievement Test Items: The Validity of Novel Types of Achievement Test Exercises in Relation to Instruction: Transferring Field Situations to Test Exercises at the Air University: Planning a Test by Reverse English: The Criterion: Why a Verbal Factor?: Vocabulary as an Arbitrary Definer of the Verbal Relations Factor: Delimiting the Verbal Factor: Verbal Factors: The Relationship Between Verbal Factor Scores and Other Variables: The Verbal Factor from the Consumer Point of View: The Value of Norms Based on a New Type of Scale Unit: Norms and the Individual Pupil: Norms and the Individual Community: Norms for Scholastic Aptitude and Achievement Tests of Independent Secondary School Pupils: Norms of Achievement by Schools: and Some Basic Problems Concerning Achievement Test Norms at the Elementary and Intermediate Grades. (MH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED181049

PROCEEDINGS

OF THE

1948 INVITATIONAL CONFERENCE ON TESTING PROBLEMS

"Validity, Norms and the Verbal Factor"

TM 009 951

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Educational Testing Service

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC).

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

EDUCATIONAL TESTING SERVICE

EDUCATIONAL TESTING SERVICE

•

BOARD OF TRUSTEES

James B. Conant, *Chairman*

Raymond B. Allen
Joseph W. Barker
Oliver C. Carmichael, *ex officio*
J. William Fulbright
Henry H. Hill
Herold E. Hunt
Katharine E. McBride
Thomas R. McConnell
Lester W. Nelson
Edward S. Noyes, *ex officio*
George F. Zook, *ex officio*

•

EXECUTIVE OFFICERS

Henry Chauncey, *President*

Richard H. Sullivan
Executive Vice-President and Treasurer

William W. Turnbull
Vice-President of Test Development, Analysis and Research

Invitational Conference on Testing Problems

October 30, 1948

JOHN C. FLANAGAN • *Chairman*

- Constructing Examinations So That They Will Be Valid Measures of Important Functions.
- Defining The Nature of the Verbal Factor As Used In Aptitude Test Batteries.
- Establishing The Type Of Norms Most Useful And Important For The Interpretation Of Achievement Test Scores.

EDUCATIONAL TESTING SERVICE

NEW YORK, N. Y. • PRINCETON, N. J. • BERKELEY, CAL.

FOREWORD

The Invitational Conference on Testing Problems which up until this year was sponsored by the American Council on Education has provided a much needed opportunity for those professionally concerned with testing to meet together and discuss their common problems. The Conference has grown from a small group to a meeting attended by over two hundred people. A debt of gratitude is owed the American Council on Education for initiating this Conference and for supporting it over the years.

With the transfer of the testing activities of the American Council on Education to the Educational Testing Service, it seemed appropriate to transfer also the sponsorship of the Invitational Conference on Testing Problems. The 1948 Conference was therefore the first held under the auspices of the Educational Testing Service. John C. Flanagan, as Chairman, planned the program and conducted the meeting which by general agreement was considered to be among the most successful in the entire series of Conferences. Since the papers presented and the discussion that followed will be of interest to a wider audience, the complete Proceedings of the Conference have been printed.

I would like to take this opportunity to express the appreciation of the Educational Testing Service to Dr. Flanagan and to all of the members of the three panels who presented papers and who took part in the discussion.

March 1949

HENRY CHAUNCEY, *President*
Educational Testing Service

PREFACE

The 1948 Invitational Conference on Testing Problems, sponsored by the Educational Testing Service, was held at the Roosevelt Hotel in New York City on October 30, 1948.

In planning the 1948 conference a number of individuals were consulted. These individuals were canvassed for suggestions as to topics which they felt represented live issues regarding testing problems, for their opinion as to which of these topics could be discussed most profitably by experts, and for suggestions as to people who, they believed, had substantial contributions to make to the discussion of the topic.

On the basis of these considerations, the final selection of topics and participants was made. It was not intended that the panel settle, or even come to general agreement on, the problems discussed. Moreover, it was planned that these discussions would not stress reports of completed work. The aim was to gather together a number of individuals who had been concerning themselves with problems related to a specific topic and to get the benefit of their observations, ideas, and tentative conclusions on the topic. It was especially hoped that such an exchange of ideas and observations would stimulate the thinking not only of the participants in the panel but also of all the participants in the conference.

The first topic, "Constructing examinations so that they will be valid measures of important functions," was selected because it seemed that the importance of the topic warranted further discussion of some of the issues raised by Dr. Rulon's remarks at the 1946 conference. Although the concept of validity in terms of content has been receiving increasing recognition, there are still many in the field who appear unwilling to accept any evidence of validity other than empirical data. The participants were selected so that their presentations would deal with three aspects of the topic. The first aspect is the one that has received the most attention from test technicians, the empirical validity problem. The second is the problem of selecting materials that will be the most valid. And the third is the broader problem of the definition of validity.

The second topic, "Defining the nature of the verbal factor as used in aptitude test batteries," appeared timely and appropriate because of the increased tendency to use specific aptitude or factor scores instead of a single general-intelligence measure, and because it is generally believed that, of the var-

ious abilities, the verbal play the largest role in educational achievement. It was hoped that the participants, all of whom have been actively concerned with this topic, would clarify the concept of verbal ability and would stimulate thought and discussion which might lead to crucial studies in this area.

The third topic, "Establishing the type of norms most useful and important for the interpretation of achievement test scores," was selected because a number of important decisions regarding types of norms will have to be made in the near future by various groups, and because the increased use of test results for guidance and counseling purposes throws a heavy burden on test norms in connection with the interpretation of test results. There are few aspects of educational and psychological measurement which are in as primitive a state as the field of test scores and norms. Although there are a few notable exceptions, current norms can be regarded as entirely inadequate for providing the information essential to the proper interpretation of test results.

The plan of the panel discussion is that each participant in turn will give an eight-minute presentation. These will be followed by a discussion among the speakers. Finally, the discussion will be opened to other conference participants.

The very fine cooperation of the panel participants is greatly appreciated. All of the eighteen busy individuals invited to take part in these panel discussions accepted, and I am sure you will find their remarks both pertinent and thought-provoking.

JOHN C. FLANAGAN, *Chairman*
1948 Conference

CONTENTS

PAGE

FOREWORD by Mr. Chauncey 5

PREFACE by Mr. Flanagan 7

PANEL I *"Constructing Examinations So That They Will Be Valid Measures of Important Functions."*

ITEM SELECTION TO MAXIMIZE TEST VALIDITY.
Harold Gulliksen, *Educational Testing Service and Princeton University* 13

CRITICISMS OF COMMONLY USED METHODS OF VALIDATING ACHIEVEMENT TEST ITEMS.
Oscar K. Buros, *Rutgers University* 18

THE VALIDITY OF NOVEL TYPES OF ACHIEVEMENT TEST EXERCISES IN RELATION TO INSTRUCTION.
Max D. Engelhart, *Chicago City Junior Colleges* 21

TRANSFERRING FIELD SITUATIONS TO TEST EXERCISES AT THE AIR UNIVERSITY.
Warren G. Findley, *Educational Testing Service* 25

PLANNING A TEST BY REVERSE ENGLISH.
Charles R. Langmuir, *Educational Testing Service and Syracuse University* 28

THE CRITERION.
Phillip J. Rulon, *Harvard University* 32

DISCUSSION 35

PANEL II *"Defining The Nature of the Verbal Factor as Used in Aptitude Test Batteries."*

WHY A VERBAL FACTOR?
Irving Lorge, *Teachers College, Columbia University* ... 45

VOCABULARY AS AN ARBITRARY DEFINER OF THE VERBAL RELATIONS FACTOR.
Edward E. Cureton, *University of Tennessee* 47

DELIMITING THE VERBAL FACTOR.
Frederick B. Davis, *George Peabody School for Teachers* . 50

CONTENTS—Continued

PAGE

VERBAL FACTORS.
Alexander G. Wesman, *The Psychological Corporation* .. 52

THE RELATIONSHIP BETWEEN VERBAL FACTOR SCORES AND OTHER VARIABLES.
William W. Turnbull, *Educational Testing Service* 54

THE VERBAL FACTOR FROM THE CONSUMER POINT OF VIEW.
Herschel T. Manuel, *University of Texas* 56

DISCUSSION 58

PANEL III *"Establishing the Type of Norms Most Useful and Important for the Interpretation of Achievement Test Scores."*

THE VALUE OF NORMS BASED ON A NEW TYPE OF SCALE UNIT.
E. F. Gardner, *Syracuse University* 67

NORMS AND THE INDIVIDUAL PUPIL.
Lee J. Cronbach, *University of Illinois* 75

NORMS AND THE INDIVIDUAL COMMUNITY.
Robert L. Thorndike, *Teachers College, Columbia University* 79

NORMS FOR SCHOLASTIC APTITUDE AND ACHIEVEMENT TESTS OF INDEPENDENT SECONDARY SCHOOL PUPILS.
Arthur E. Traxler, *Educational Records Bureau* 81

NORMS OF ACHIEVEMENT BY SCHOOLS.
E. F. Lindquist, *State University of Iowa* 95

SOME BASIC PROBLEMS CONCERNING ACHIEVEMENT TEST NORMS AT THE ELEMENTARY AND INTERMEDIATE GRADES.
Walter N. Durost, *Boston University* 98

DISCUSSION 102

APPENDIX 114

PANEL I

Constructing Examinations So That They Will Be Valid Measures of Important Functions

Item Selection to Maximize Test Validity

by HAROLD GULLIKSEN

I. *Introduction.*

I will discuss the following problem: Given a set of M items, to select a subset of K items that will constitute the most valid K-item test obtainable from the original set of items. In order to do this it is necessary to write the test validity coefficient as a function of certain item parameters.

II. *Test-criterion correlation as a function of item parameters.*

The correlation between a criterion (c) and a test (t) composed of items (i) is given by:

$$R_{ct} = \frac{\sum_{i=1}^{K} r_{ic}\sigma_i}{\sum_{i=1}^{K} r_{it}\sigma_i} = \frac{\overline{r_{ic}\sigma_i}}{\overline{r_{it}\sigma_i}}$$

where,

r_{ic} is the point biserial item-criterion correlation

r_{it} is the point biserial item-test correlation

σ_i is the item standard deviation, $\sqrt{p_i\ (1-p_i)}$, where p_i is the proportion of persons answering item (i) correctly.

The bar over a term is used to designate the average. Since the same number of items is involved in both the numerator and denominator, the ratio of the averages is the same as the ratio of the sums.

It should be noted that the only assumption involved in this equation is the assumption that $\Sigma\widetilde{r_{it}\sigma_i}$ will be nearly equal to $\Sigma r_{ie}\sigma_i$, where r_{ie} is the point biserial correlation between the item and the test composed of the selected subset of items, while r_{it} is the point biserial correlation between the item and the *total* test which is composed of both the selected and the rejected items.

III. *A graphic selection procedure.*

No analytically rigorous method for using this equation to select the best set of items has yet been devised. However, the following graphic procedure gives results which are probably not far from those which would be given by the best method.

Represent each item as a point on a scatter plot with the item *validity index*, $r_{ic}\sigma_i$, as the ordinate, and the item *reliability index*, given by $r_{it}\sigma_i$, as the abscissa.

In general the points will fall to the right of the y-axis and between the lines $y = x$ (line OA)

and $y = -x$ (line OB). That is to say, negative reliability indices will be rare, and it will also be rare to have the validity index greater in absolute value than the reliability index.

Any set of selected items will have a centroid. The y-projection of this centroid will be the average validity index for the set, and the x-projection will be the average reliability index for the set. The ratio of these two indices will be the ordinate of the point divided by the abscissa.

In Figure 1, the cross indicated by "C" represents the centroid of the items in the shaded area. The tangent of the angle COX is the validity of the test composed of the set of items whose centroid is C. A subset of K items which is reasonably close to the best subset can be obtained by pivoting a ruler at the origin (O) and swinging it down until the desired number of items are above the ruler. The centroid (C) of this subset can then be found, and the tangent of the angle COX is the validity of the test made up of this subset of items.

IV. *Illustration of use of method.*

The Mechanical Comprehension section of the Officer Qualification Test was developed first solely on the basis of information about the item-test correlation. Later, when grades became available, we applied the method outlined above to see how much difference it would have made in test validity.

A scatter plot of item-validity against item-reliability index was made, and the sixty (60) best of the 104 items available were divided into two thirty-item tests. For each of these tests twenty items were identical with the items formerly chosen solely on the basis of reliability, and ten items were replacements.

The two forms chosen on the basis of reliability information only had validity coefficients of .35 and .38. The new tests chosen on the basis of both validity and reliability indices had validity coefficients of .46 and .48, respectively. (The validity for a new sample would doubtless be lower than .46.)

V. *A maximizing method which includes considerations of sampling.*

It should be noted that the method suggested above gives no indication of when to stop adding items to a test.

It is possible to select the best thirty, or the best thirty-five items out of a set of 70 items, but the method does not tell us whether it is better to use the thirty-item test, or the thirty-five item test. It will always tell us that the expected validity of the thirty-five-item test is less than that of the thirty-item test.

It may be, however, that the lower bound for the validity coeffi-

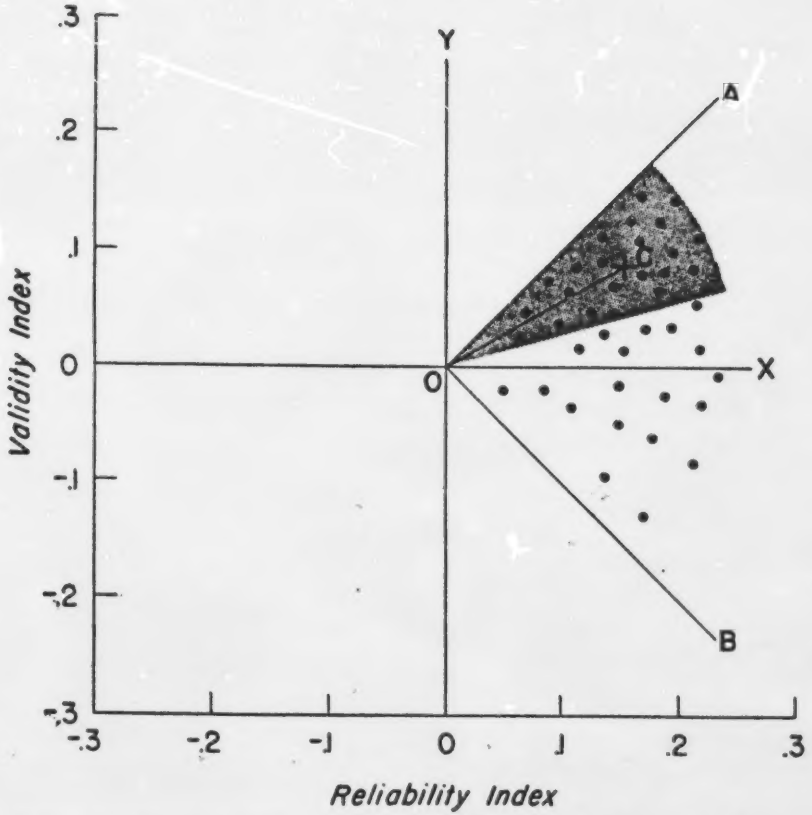

FIGURE 1. *Illustrating the selection of the most valid subset of K items.*

cient will have a maximum in some instances. In Figure 2, C_1 represents the centroid for the best two items, and C_2 represents the centroid for the best fifty items, say. The estimated validity for the two-item test is clearly higher than the estimated validity for the fifty-item test. However, if the sampling variance of C_1 could be figured, it would probably be very large (as indicated by the circle about C_1) so that the lower bound for the validity would be indicated by something like the tangent of L_1OX. On the other hand, the sampling variation for C_2 would be small, so that the lower bound for the validity would be indicated by something like the tangent of L_2OX, which is larger than the tangent of L_1OX despite the fact that the tangent of C_2OX is smaller than that of C_1OX. That is to say, the lower bound for validity may have a maximum, while the centroid will not. Thus, if one wishes to be fairly certain of obtaining a valid test, such a maximizing procedure may indicate not only which are the best items, but also how many items it would be desirable to use to be relatively certain of having a test that would not drop below a specified validity.

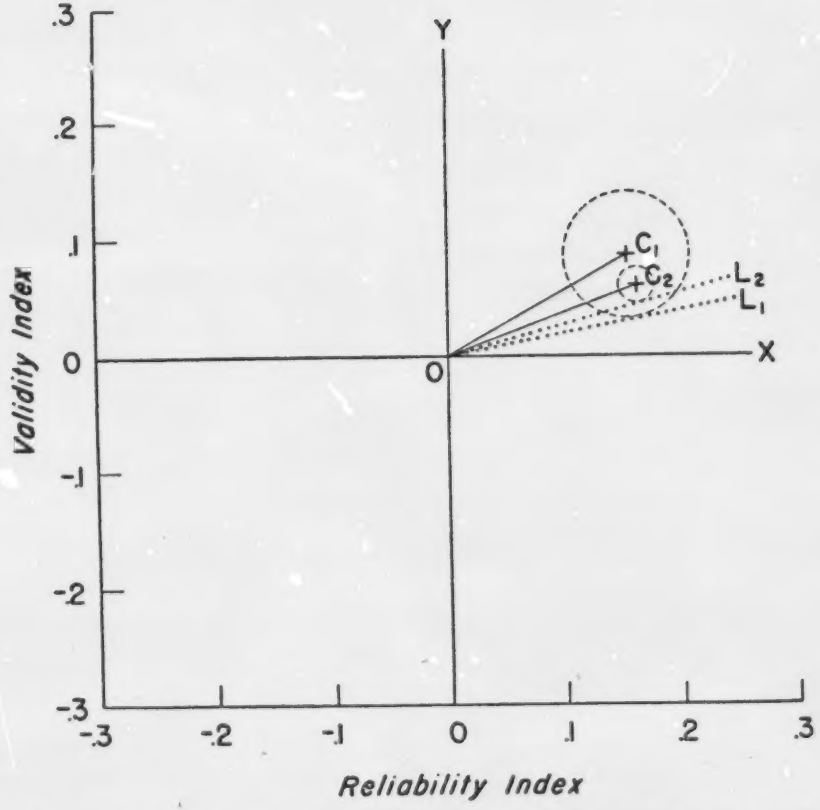

FIGURE 2. *Illustrating the possibility of a maximum for the lower bound of the validity coefficient.*

Criticisms of Commonly Used Methods of Validating Achievement Test Items

by OSCAR K. BUROS

With the hope of provoking discussion and thought regarding the adequacy of many of our practices in the construction of achievement test items, I should like to make several suggestions. Since my time is very short, I shall present each suggestion rather briefly and somewhat dogmatically. If you wish me to elaborate upon any of these points during the discussion period, I shall be glad to do so.

1. *The practice of validating achievement test items by selecting only those items which are answered correctly by an increasing percentage of students in successive grades should be abandoned.* This procedure was, to my knowledge, first used in the construction of a school achievement test by three of our most distinguished testing specialists—Truman L. Kelley, the late Giles M. Ruch, and Lewis M. Terman—when they prepared the Stanford Achievement Test over twenty-five years ago. Practically all achievement batteries today use the same procedure of item validation; e.g., this was done in the recently revised *Metropolitan Achievement Tests* and the recently published *Coordinated Scales of Attainment.* This criterion begs the question and results in emasculated tests of questionable validity. No test validated in such a manner can be used to determine differences in grade achievement. The tests which result exaggerate differences in grade achievements, conceal areas in which little or no learning has taken place, and cover up areas in which there have been actual decreases in student attainments.

2. *The practice of validating achievement test items by selecting only those items which are most closely correlated with the total score should be discontinued.* This criterion also begs the question and results in emasculated tests of questionable validity. No test validated in such a manner can be used to determine differences in the performance of the so-called poor and good students. The tests which result exaggerate differences in student achievements, give a lopsided emphasis to those factors represented most heavily in the total score, and fail to give information about growth or lack of growth toward certain desired learning outcomes.

3. *The practice of considering "measurement" and "differentiation" as synonymous terms should be abandoned.* A test may be an excellent measuring instrument and yet not differentiate among individuals within a given group. Our objective should be primarily to measure with a given degree of accuracy--differentiation may or may not follow.

4. *The practice of discarding items merely because they are passed or failed by all examinees in a tryout group should be abandoned.* If the items have curricular validity, they should be retained both for purposes of measurement and for purposes of motivation. Under no circumstances should items be omitted solely on the grounds that only chance scores were obtained in a tryout group.

5. *We should give more attention to subjective procedures for evaluating test items.* In most cases, the best way to determine item validity is to study carefully the nature of the response which the item elicits and then to decide whether the response would give any useful evidence concerning growth or lack of growth toward the particular objectives which the test is designed to measure. Unfortunately, this procedure will call for much more work and ability than our present cut-and-dried methods of item validation by the statistical methods which I mentioned earlier. Additional procedures which might well be used to validate individual test items would be the use of a free response or short essay question covering the same ground and the use of individual interviews in which the examinee is probed to ascertain his understanding of the concept covered by the question. It also should be remembered that checks on the validity of test items must be repeated at intervals in order to determine whether changes in the learning situation may have diminished or destroyed the original validity of the test. For example, coaching exercises designed to help students pass reading tests have certainly vitiated to a large extent the validity of many of our reading tests.

6. *We should enlist the cooperation of specialists in subject matter, teaching, and curriculum construction.* This cooperation must not be of the kind which has characterized most of the test construction work to date. At present there is too much of a tendency to enlist the support of only those subject-matter and curriculum specialists who will go along with the test technician in his erudite methods of statistical validation. The specialists in other areas should be invited to cooperate as coequals and not be intimidated by technical jargon which makes no sense to them.

7. *Test authors and publishers should be impelled to present a great deal more data than they now do concerning the methods which*

they used to secure a valid test. Vague statements such as, "The customary methods of curricular and statistical validation of test items have been used to ensure the best possible test," are useless and misleading. If an analysis of textbooks was made, a detailed description of the analysis and the results should be reported in order that others may judge the study and the use made of it. In an equally as detailed manner, all other aspects of test construction should be described. I should like to suggest that test authors and publishers go further. Let them point out common objectives not measured by their tests; let them suggest ways in which the tests may be supplemented locally in order to have a reasonably well-rounded evaluation program. I am confident that the publication of much more detailed data regarding test validity will result in criticism which will force test makers to construct better tests.

The Validity of Novel Types of Achievement Test Exercises in Relation to Instruction

by MAX D. ENGELHART

In recent years there have been numerous efforts to produce objective achievement test exercises which will measure more than the information or knowledge possessed by the student. Examples of such exercises are to be found in certain of the tests used in the Eight Year Study of the Progressive Education Association and in the Cooperative Study of General Education. Such exercises are also to be found in some of the subject tests prepared by the Examinations Staff of the United States Armed Forces Institute, in comprehensive and other examinations constructed by boards of examiners of higher institutions, and to a more limited extent in newer standardized achievement examinations.

Only brief consideration can be given at this time to the various types of objective exercises which may be useful in measuring more than memory of facts. A true-false or multiple-answer exercise may accomplish this end if the content of the exercise is to some extent novel to the student. The item stem of a multiple-answer exercise may present the student with a situation requiring an application of information which has not been mentioned in instruction. Suppose, for example, that a student has learned that gases and liquids are fluids, that gases expand on heating, and that, according to Archimedes' principle, a body is buoyed up by a force equal to the weight of the fluid displaced. Then, if the instructor has not made explicit this application, more than sheer memory of these facts should be required in successfully responding to an objective exercise based on the question, "Will a dirigible balloon rise more rapidly in warm air or in cold air?" Along with novel content of the item stem, it seems essential that the various alternatives or distracters be sufficiently plausible as to promote discrimination between them.

Other types of objective exercises which are relatively simple in form include lists of items to be classified according to certain categories. Such exercises may require

more than memory of facts, if the student needs to range over the various bits of information he has acquired and make discriminations he has not previously been required to make. For example, suppose that instruction has first been concerned with various phenomena which are characteristic of sound and that later instruction has dealt with various phenomena characteristic of light. Let us also suppose that the instructor has not made comparisons between these two groups of phenomena. Then it is probable that the student has to do more than just remember when classifying such items as "refraction," "resonance," "diffraction," and "will travel in a vacuum" according to the categories "true of sound," "true of light," "true of both sound and light," and "true of neither sound nor light."

Paired items of certain types may force students to make comparisons. The student is to mark "A" if the first item is greater than the second, "B" if the second item is greater than the first, and "C" if the two items represent essentially equal magnitudes. For example:

Extent to which heat regulation of the body is due to the control of heat production.

Extent to which heat regulation of the body is due to the control of heat loss.

Paired items may also be written to measure the student's ability to deal with causal or correlated relationships. The student marks "A" if increase in the thing described by the first item is accompanied by increase in the second, "B" if increase in the thing described by the first item is accompanied by decrease in the second, and "C" if neither of the above relationships exists. As an example, consider the following:

Extent to which President Truman advocates civil rights.

Number of votes he will receive in the South on election day.

Numerous thought-provoking exercises have been prepared in which a student is presented with information or data in a paragraph or longer selection, or in the form of a graph. One of the tests of the Progressive Eight Year Study, Test 2.51, Interpretation of Data, presents data in these ways and requires the student to classify items pertaining to selections or graphs according to the categories, "true," "probably true," "false," "probably false," and "evidence is not sufficient to indicate that there is any degree of truth or falsity in the statement." One can obtain from a test of this kind a general accuracy score and also scores reputedly measures of the extent to which a student tends to over-generalize, is over-cautious in interpreting data, or makes crude errors of judgment. A recent examination in biological science constructed at the University of Chicago contains an exercise of somewhat similar nature. This

exercise begins with a series of numbered observations. This is followed by a series of "conclusions" which are to be classified "A" if warranted by the data, "B" if contradicted by the data, and "C" if the conclusion goes beyond the data. Each of the conclusions is also to be related to the particular observations which are relevant to it by appropriately marking the answer sheet.

Other examples of objective exercises designed for the purpose of measuring more than the retention of information could be given. It is possible, however, that a sufficient variety have been mentioned to form a basis for discussion of the importance of relating achievement exercises to the objectives of instruction. Reference has just been made to the biological science exercise prepared at the University of Chicago. The speaker has no reason to question the validity of this exercise so far as biological science instruction at the University of Chicago is concerned. I feel very sure, however, that this exercise would lose validity if included in one of the biological science comprehensive examinations of the Chicago City Junior College. Many of our students would not know what it is to go beyond the data. It is even more certain that the phraseology of certain of the items would baffle them. These items contain such phrases as "a sufficient cause of . . . .," "a necessary cause of . . . .," and "a cause of . . . . which is both necessary and sufficient . . . . ."

In writing or selecting exercises whose function it is to measure the retention of information it is relatively easy to determine face validity. It is much more difficult to determine the extent to which exercises may be novel in form or content without being so novel that students are unable to do the hoped for thinking with reasonable success. One solution to this problem is to analyse exercises with respect to the concepts not made explicit in the exercises, but which are necessary to their solution. In the case of the exercise concerned with the question, "Will a dirigible balloon rise more rapidly in warm air or in cold air?", one would need to determine whether or not instruction has given the students an opportunity to learn that gases and liquids are fluids, that gases expand when heated, that air is a gas, and that according to Archimedes' principle a body is buoyed up by a force equal to the weight of the fluid displaced. In the case of exercises involving the interpretation of data, analysis of the exercises is necessary to discover whether or not students are likely to be prevented from doing the necessary thinking because of unfamiliar terms in either the items or the directions, or because they have never had experience in doing the kind of thinking required by the exercise. If instruction has merely been the imparting of

knowledge or information, the use of such exercises cannot be justified as a means of evaluating the attainment of the objectives of the instruction.

It is widely accepted that the formulation of the objectives of instruction is the function of the instructor rather than of the examiner. While this principle is fundamentally sound, it does not seem inappropriate for an examiner to seek clarification of instructional objectives and to use opportunities which arise in test construction and evaluation as means of increasing the scope of desirable educational objectives. Many instructors recognize the importance of objectives which are not the real objectives of their instruction. There are few teachers of science who do not consider the ability to do critical thinking in their field an important objective. When teachers participate in the construction or evaluation of exercises of the types described above their real objectives become more apparent to them and they may themselves become more aware of the needs for both extending the scope of their real objectives and of modifying instruction accordingly. On one occasion the speaker constructed an exercise organized in such a way as to conform with the usual steps of scientific method. It began with the statement of an hypothesis; data were then given in a paragraph. This was followed by a number of inferences to be evaluated in terms of the data; and finally came the selection of the conclusion best justified by the data. When this exercise was discussed with the chairman of the science department concerned with the examination in which the exercise might appear, his reaction was, "We have always stated in the preface of our text that scientific method is one of the major objectives of our instruction. This is the first time anything has ever been done about it."

Much research needs to be done by psychologists and test experts to determine the validity of the various types of novel achievement exercises. There should be more studies in which achievement tests and psychological tests given to the same populations of students are subjected to factor analysis. There should be more attention given to the logic of achievement exercises and to the problems of semantics involved. But it should not be forgotten that the writing of exercises is an art requiring imagination and inspiration and finally, that, if the exercises are to be valid measures of the objectives of instruction, teacher participation in construction or evaluation, while not sufficient, is certainly necessary. It is only through this means that achievement testing can realize its possibilities in the improvement of measurement and of instruction.

Transferring Field Situations to Test Exercises at the Air University

by WARREN G. FINDLEY

The ultimate criterion of successful instruction is on-the-job performance. The face validity of any evaluation procedure, therefore, increases as it approaches the job or field situation. In the evaluation of learning in advanced professional training, where high professional competence is the expected outcome of the training, close approximation to a realistic field situation is especially essential. For it is virtually impossible later to secure comparable field-situation data on the students, against which to establish empirical validity. Advanced training of military officers, as at the Air University, presents this problem in acute form.

This approach to realism in the evaluation or testing procedure needs to be along two main dimensions: (1) the scope and nature of the test situations set, and (2) the extent to which the situations represent the variety of problems inherent in the field situation.

Under scope and nature of situations, the requirements might be enumerated to include (a) participation in group solution of problems rather than individual solution, if the field situation inherently involves a group solution, (b) the carrying through of large units of thinking rather than parts of such units, when in a field situation a crucial element is the ability to handle large units of thinking, and (c) presentation of realistic elements of data, combined in such a way as to retain realism in the total situation presented.

The second main emphasis, on the need for representing the variety of problems inherent in the field situation, is closely related to the last requirement. A single problem usually cannot be made to test all the important understandings about a subject and remain real. Conversely, however real some problems may be, they are representative of only a small fraction of the great variety of situations a professional worker, like a field-grade officer, may be expected to face. As a consequence, if a limited time is available for evaluation, either problems must be reduced in scope to permit sampling by the use of several situations, or necessarily long problem units must be supplemented by miscellaneous items of lesser face validity.

In the Air University, group solution of problems constitutes an increasing proportion of activity as one passes from the Air Tactical School (for junior officers), through the Air Command and Staff School (for field-grade officers), to the Air War College (for senior officers). The seminars and staff problems involve group solutions to problems in which each officer is graded by his colleagues and/or an observer on descriptive trait-scales for his performance in group work.

Different approaches have been used in different schools to attain the combination of proper scope, realism, and variety in individual problems. In some cases the outcomes expected are pencil-and-paper outcomes. The writing of operations orders in conformity with regulations governing military correspondence is a case in point. Other exercises represent a small shift, such as the inspection of flight records, in which a representative set of such records is reproduced and students are required to indicate faults and corrections in a quite objective fashion.

Where the field situation is broader, as in the planning of a strategic or tactical bombing mission, or in the handling of personnel problems, the approach has been to describe such field situations verbally and pictorially and to ask the student to respond with a free answer, or objective indication of choice, to the full situation. The planning of a mission generally permits sufficiently full coverage of outcomes of instruction about such planning to stand as a single large problem. The variety of personnel problems dictates a series of briefly described situations, the solution to each of which is to be given in a short paragraph.

At the Air Command and Staff School (for field-grade officers) the tendency is to feel that only full three-hour problems can do justice to realism in most situations. If, then, the realistic problem cannot test all the principles covered in a subject, the tendency is to supplement the problem by miscellaneous objective items covering the points not illustrated.

At the Air Tactical School (for junior officers) the tendency has been to feel that mastery of the job requirements of such officers can generally be well tested by a series of short problem-situations, each accompanied by two or three objective questions. These situational problems are then supplemented by miscellaneous objective items if all points cannot be readily reduced to situational testing.

One device that has proved helpful in objectifying the scoring of full-scale problems is to divide the problem into its major phases. After allowing sufficient time for the first phase, an intermission is given during which papers on the first phase are collected; an official

solution of the first phase is then distributed and the solution of the second phase is required. Subsequent intermissions are granted at the end of each additional phase, at which times similar procedures are followed. As a result, all students are set the same task in each phase, problem scope and continuity are maintained at a high level, and each portion of the test is available for separate scoring.

Planning a Test By Reverse English

by CHARLES R. LANGMUIR

Constructing tests so that they will be valid measures of important functions is something test-makers are trying to do all the time. One very common procedure is to start with some general label or name for some important function, for example, reading, or quantitative ability, or space relations, or reasoning ability. After a quick, or maybe not so quick, analysis of what people mean by such labels, the test-maker collects some test items that seem to be loaded or possibly saturated in some way with the important functions characterized by the label. He tries the test on some students and computes means, standard deviations, and especially a table of percentiles.

Another test-constructor is likely to start out with some items for which he has a particular fancy. He chooses a nice name for the collection of test items. He finds a sample of some population and he computes means, standard deviations, and especially a table of percentiles. In this case, the test-constructor chooses a criterion which, regardless of his descriptive label, is actually defined by the items he started with. The procedure has the merit of definitely stating that the criterion is the sum of all the items. A valid test of an important ability may be the reward, provided a lot of other people who don't care much about the details of test construction also take a particular fancy to the items. Here, of course, I mean that they take a particular fancy to the particular items and not just a liking for the fancy name.

Both of these methods are relatively easy. It is just very unfortunate that it is usually impossible for a test-user to find out to what "degree the test actually measures what its author purports it to measure," and whether the function measured is after all of any real importance.

There is a third method which is not so easy but which has the advantage that the test-user does not require any secret knowledge of test magic in order to know what he has got. By this method the test-constructor will start with a score. He will assume a score, just one score. He will then write a statement of what he wants that score to mean, and I am quite sure it will not be a statement that says that this score is the fourteenth percentile in a distribution of several thousand scores. Preferably, indeed essentially, the statement will

have unambiguous meaning to a second person whose experience is unadulterated by any interest in test-making. When the test-maker has written out satisfactory statements of the meanings which he desires to attach to the scores, it is very likely that some good suggestions about the form, nature, and details of the desired test will be obvious.

The possibility of constructing valid tests by this "reverse English" is nicely illustrated in a by-product of a recently completed project. Kurtz and Seashore had a problem of devising a method of selecting enlisted men who would be successful in radio operator school. Naturally, the first thing required was a criterion measure of success in school. Naturally, school grades, instructor ratings, etc., were considered and rejected. Naturally, analysis of the problem indicated a test of ability to do the things the school was set up to teach. After due attention to the problem of sampling letters, digits, and special symbols, and after proper consideration of comparability among forms, methods of standardizing administration, and problems of scoring, some achievement tests were constructed. The achievement tests were used as the criterion of successful learning and a good method of identifying men who would succeed in learning Morse Code was determined. The selection materials were produced, reports were written, the Navy was happy, the investigators went off to other work, and everybody was satisfied except Langmuir.

From my point of view it was unfortunate that the project was terminated, because a little additional work would have carried the achievement test which was only a by-product of the job to an elegant completion. The kind of meaningful validity statements that I suggest as the best way to start building tests were implicit in the development. If the work had continued, the type of validity statement I have in mind would have been made explicit. The project was officially concerned only with the selection of men for assignment to school. If it had been concerned with the classification of men for assignment to more crucial responsibilities, the significance of the achievement test development probably would have been recognized.

A specific example may clarify the notion. Let us start with a score on a test of ability to receive Morse Code and see what we get into when we try to write a statement of meaning which is sensible to a second person who knows nothing about tests.

Assume a score, $X = 18$. As a first effort we say that the score 18 identifies a man who can copy Morse Code at the rate of 18 words per minute. Obviously we need to add to this initial statement something

about the errors that such a man will make. The score of X = 18 identifies a man who can copy Morse Code at the rate of 18 words per minute with an expectation of two errors in 1000 characters. We are free, of course, to write our statement with some other expectation of error.

The statement seems to be reasonably complete in the sense that it is definite with respect to the two most important elements of performance, namely speed and accuracy. But it is not complete in one vital sense. It does not tell us what we may expect such a man to accomplish on easier and harder tasks of code reception. Specifically, we wish to know the expectation of errors at 12, 14, and 16 words a minute and for the harder tasks, 20, 22, and 24 words per minute.

If the relative frequency of errors on easier and harder tasks is substantially the same for a normal range of operators, it follows automatically that our statement of meaning of the score must include a reference to data describing the error performance of radio men on tasks both less and more difficult than their standard score level. If, however, as seems likely, operators who score 18 vary a great deal with respect to their errors at other speeds, a second score is required. The probable nature of the data suggests to me a subscript, say alpha, to the score which will identify a certain characteristic curve or table of data which describes the expected performance of the man.

Upon further consideration of what we want a score of 18 sub alpha to mean, it is likely that we will want to know what a man who scores 18 sub alpha may be expected to accomplish when he receives code under conditions of heavy static or in noisy radio rooms where there is a good deal of distraction. If operators vary a great deal in their ability to adjust to deviate conditions, a third score possibly in the form of a superscript is indicated.

Here we have an example of what happens when we try to state the meaning we wish to attach to a test score in advance of having the test. We have discovered in the statements of meaning not only some notion of the character of the test, but we have also discovered that the form of score we selected is not sufficient. In short we have to add something in order to make the reported score carry all the meaning we wish to have it carry.

If someone should suggest that I have dug up an old chestnut in urging everybody to work harder on the criterion, I should have to agree that that is just what I have said. But I suggest a procedure here for working on the problem of the criterion that may possibly yield more results than we are accustomed to getting. If we start with a score and write a statement of meaning, I believe that we will

more quickly and more precisely determine the elements of abilities which are usefully important.

The power of the method lies not in the art of test construction nor in the statistical manipulation of variance, covariance, and percentile data. The importance of writing a statement of the desired meaning of a single test score is simply that the existence of such a statement makes explicit the kind of data which are required to get the score. If we can describe exactly the kind of data required we can very likely think of a way to get them.

If we should find that our statements specify data which describe the performance of people on tasks other than tests, and if we should wind up with tests which turn out to be observations of performance very much like work performance, we should not be very surprised.

The Criterion

by PHILLIP J. RULON

In the course of a normal life span of a human being living in our culture, he carries out a tremendous number of activities, using a vast variety of things under a very wide diversity of circumstances. This is a rather trite remark. Some of these actions are the concern of educators and psychologists and are studied by them, more or less exhaustively. We find in such study that almost no single action by a human being is regarded as important in itself. Now, some few single actions, like murder, embezzlement, and so on, are the proper concern of psychiatrists and clinical psychologists, and the prediction of these single actions is, of course, very important. But educational psychologists are more interested in series of actions. We might say we are interested in the single act of marriage, but actually we are more concerned with whom to marry and how to act after marriage than we are with the question of whether or not one ought to marry.

One of the hardest things to get agreement on is what actions and conditions shall be considered important. In particular, it is very hard to get agreement on what actions and conditions are considered *desirable*. Obviously, the only ultimate criterion of a successful life would be—or if not obviously, then presumably—for the individual to die happy after a long and happy life, with acceptable reasons for the happiness. But we can't agree on what we mean by happiness; we can't wait for the people to die, and, even if we could, we couldn't make very many exhaustive studies of the people after they had done so.

So, in trying to agree on what actions and conditions are desirable, we must make value judgments of less than ultimate values. We may as well face this fact. These cannot be ultimate value judgments. Deciding what is an ultimate good has occupied philosophers and mystics for centuries, and, if we insist on an agreement on ultimate good, we shall be concerned for centuries, and we won't get any measurement work done.

Thus, all of our value judgments are short of ultimate, and we may as well recognize it. In particular, we may as well stop criticizing people who try to predict desirable behavior when the behavior is less than ultimately desirable. For ex-

ample, success in school. I say in essence that there is nothing the matter with trying to predict success in school. It isn't ultimate, but the same criticism can be leveled against almost any more nearly ultimate criterion. We can try to get as nearly an ultimate criterion as possible, and stop at the point of diminishing returns.

Since no single action is regarded as very important, we ask about consistent or habitual performance. In studying the actions of people, we discover that consistencies or uniformities do appear. We also discover with respect to certain kinds of behavior that a change in conditions is followed by a change in the behavior. For example, the offering of a reward of some kind may be followed by a marked change of behavior. Miss Jones may appear very promptly at the office in the morning when pay deductions are introduced as penalties for tardiness, or when her office group is offered a free Christmas trip if everybody is on time in the mornings. We call such readily modifiable behavior voluntary. A consistent performance of such activity is called habit. Note that the habit does not cause or explain the behavior. The habit *is* the consistent behavior.

We may find that certain other kinds of behavior are *not* markedly changed by the offering of rewards. Miss Jones may do no better in explaining or acting upon (that is, comprehending) complicated instructions when a free trip is offered than when no reward is offered. We do not have a word for such activities, since involuntary behavior has come to mean something else, but we do have a word for the consistent carrying out of such not readily-modifiable behavior. We call the consistent performance ability. Note that ability does not cause nor explain the behavior. Ability *is* consistent performance. If a person performs with less than perfect uniformity, he will sometimes perform above his ability level and sometimes below, and there may be no significance at all to an individual's performing below his ability level.

Going back to voluntary behavior again, we note that the conditions under which the behavior changes are called motivation. To improve motivation, we make a study of the conditions under which behavior is changed in the direction we desire. We also, as human beings sometimes do, confusingly use the term "motivation" to cover up our ignorance of conditions. Thus, when an individual does not behave the way we wish, we may retreat into the meaningless statement that his motivation is weak.

We have also used the word "ability" to cover up our ignorance. When an individual does not perform a task we want him to, we may say he has no ability, or little ability.

The only way to avoid confusion in thinking about these things, and

hence to avoid silly policies and silly decisions on the part of educators, is to employ behavioristic definitions of the things we are trying to observe, and also behavioristic definitions of the things we are trying to accomplish. We need such an approach to interest, to aptitude, capacity, skill, proficiency, achievement, and—heaven knows—we need such a concept of measurement itself. Having agreed upon what we mean by these things in human adult behavior, we can next attempt to agree in the case of school children, and then we can observe the things, and, reporting numerically upon our observations, we may say we have "measured" them.

DISCUSSION

Participants:

WARREN G. FINDLEY, PHILLIP J. RULON, JOHN C. FLANAGAN, HAROLD GULLIKSEN, CHARLES R. LANGMUIR, OSCAR K. BUROS, WALTER N. DUROST, HENRY S. DYER, EUGENE D. CARSTATER, J. WAYNE WRIGHTSTONE.

CHAIRMAN FLANAGAN: I would like at this time to throw the general problem open to discussion by the panel. Certainly we have had a good deal of agreement on the importance of what is being measured and the importance of some of the logical problems in getting the criterion.

I believe that the disagreement this group has expressed with current practice is that while everybody is concerned about the criterion, not very many people have spent very much time on it, and that, instead of spending 95 per cent of our time going through a certain set of operations to get ourselves items and 5 per cent with the criterion, or with what those items are measuring, perhaps even the reverse percentage for a while might pay off in results.

Would some panel member like to make some comments about other people's presentations?

DR. FINDLEY: I would like to make a comment on one of Dr. Buros' points that we should distinguish between differentiation and measurement, his point being, as I remember it, that you might have 100 per cent achievement on some things, which would reflect the fact of the situation, and that differentiation forced upon such a situation would be out of place. I think there are some types of material in which 100 per cent mastery may be expected; but it has been our observation, in making questions that call for the application of understanding to realistic problems, that unless the items are of the sort that result in a differentiation of the group the items are themselves faulty. By setting problems that require application we have automatically forced upon people problems that will result in differentiation, which is a part of the measurement task.

DR. RULON: Mr. Chairman, I suppose there must be some resolution to this issue; certainly I have to go along with Buros. A couple of examples, I think, will indicate why there is an issue, and I will leave to somebody else the resolution of it.

When Dr. Ben D. Wood was first starting his national teacher examination program, he went into the studies of health of teachers as a predictor of teaching success and found that health had no validity for predicting teacher success; it didn't have anything to do with teaching success. So then the next question was, should you disregard health altogether in hiring a teacher? Well, most superintendents think that they want to have healthy teachers, and presumably an absolute measure of the health of the teachers is a descriptive report on their health, as Mr. Buros put it, regardless of the fact that such a health statement makes no differentiation among the teachers.

A still sillier case happened in the beginning days of Dr. Flanagan's Air Force program. Of course you all know Air Force pilots have to have 20/20 vision. That isn't true in the British Air Force, and it certainly wasn't true in the Japanese Air Force, but it was true in our Air Force. Some measurement expert came along and found that visual acuity had no value whatever in distinguishing a good pilot from a bad pilot, and therefore proposed that visual acuity be dropped. How would you like to fly around with a blind pilot?

CHAIRMAN FLANAGAN: Does anybody want to answer that?

DR. GULLIKSEN: I don't know that I will answer that. I think you might drop the requirements below 20/20, but not quite as far, possibly, as was suggested.

I would like to comment on and extend in particular some of Engelhart's remarks. It seems to me very important to emphasize that the examination should be intimately tied up with the teaching. Consider, for example, his first illustration, "Will a dirigible go up faster in warm or cool air?"; that is a reasoning problem only if the students have not been told the answer last month and so remember it. It seems to me that the next important step toward actually doing something about this criterion that we have been discussing is to tie examination construction far more closely to the teaching. I think that statewide, or nationwide, testing programs are very good for certain purposes, but such programs have not achieved the integration of examining and teaching which is necessary. The problem is how to tie a fairly good examining program closely to classroom teaching. The teacher should set up the questions, bearing in mind exactly what has been taught and exactly what has been omitted. The examination is set up with the notion that the curriculum has done this and has omitted that and now we will see what the students know. After the examination the teacher may alter the curriculum a little and try the same or a slightly different examination.

It seems to me that such procedures are possible only on a small scale, and then one usually runs into a financing difficulty. People trained in examination techniques, as well as people trained in subject-matter techniques, aren't available to work together on the problem. How can such an integration of examining and teaching be financed? Such an integration of teaching and testing seems to me to be the first step in improving the criterion against which tests are measured.

MR. LANGMUIR: I would like to ask Mr. Buros to elaborate a little bit on his notion that test constructors should actually abandon statistical criteria in the selection of items. If I understood Buros correctly, he said that we should actually disregard observations of difficulty differences in grade level. If we give an item to a fourth-grade class and 25 per cent of the children in the fourth grade answer this particular item correctly, and we try the same item in the fifth grade, and 25 per cent of the fifth grade answer it correctly, and we give the same item on through the school with the same result, I think we shall very likely not regard that item as suitable for use in a test for the purpose of differentiating or classifying pupils, or for giving information to the instructor, to the principal, to a superintendent, to a survey committee, or, in fact, to anybody about the individual performances and changes of performance that take place in the school.

Clearly, if we abandon statistical criteria entirely, we are going to have collections of items about which we know nothing. Indeed, the thought was made so strongly that I was tempted to think that perhaps Buros felt that we should have tests made up of items which definitely do not differentiate among pupils.

DR. BUROS: I mean exactly what I said; I think that that particular criterion for selection of items for achievement tests should be abandoned. I first became aware of that difficulty many years ago when I was an elementary school principal. In my school we were teaching American history in the fifth grade and European background in the sixth grade. We were making use of the Stanford Achievement Test. One of the social studies sections of that test was devoted almost entirely to American history, and I became very much interested to find that my children who had been away from American history for one year did better on this test than the fifth-grade children, who were just completing American history. And then, when I started to study the manual more carefully, I realized that that was bound to happen; if my elementary-school children were similar to the standardization group which Kelley, Ruch, and Terman had used in constructing the test, naturally we had to find exactly what we did

find, since they had assumed that sixth-grade children must know more American history than fifth-grade children, and since there were very few items on European background in that test.

I might mention another instance. About twelve or thirteen years ago you may remember that the American Historical Commission was making a study of the teaching of the social studies in the schools, and test construction played a very important part. We had Historian Krey of the University of Minnesota and Statistician Kelley of Harvard University cooperating in this affair. Well, we had various subject-matter specialists constructing these tests under the guidance of Dr. Kelley.

The main criterion which Dr. Kelley set up for determining whether items were good, bad, or whatnot, was this increasing percentage of correct responses as you went through from grade to grade. Well, the report ended up in a very interesting manner. The historians were very much puzzled. Krey made some such statement in his report as this: He said, "We subject-matter specialists were very puzzled by the procedures used in constructing these examinations because we saw many items which seemed the very best items to us, and which measured the most important concepts, rated as inadequate and invalid and rejected. We saw other items which seemed very poor, but, since they met the statistical criteria set up by the director of the study, they held them, as I think almost always happens when the subject-matter specialists cooperate with the test technicians. With these misgivings, but determined to follow out the dictates of scientific method, we decided to go along with the test experts."

Well, now, that whole story was just one series of disappointments in this particular case. Let us just stop to reason about it very simply. If you assume that sixth graders do know more than fifth graders, and you construct tests which are based upon that assumption, how can you find anything but what you have assumed? You have a case of circular reasoning there. You are begging the very question. These tests are not measures of achievement.

I suppose what has happened there is that that particular criterion is not a bad criterion for a selection of items for an intelligence test, and that it wasn't a particularly bad criterion for the categorization of people in the Army trade test developed in the First World War; and since it worked out very nicely in those two situations, it was naturally picked up and used in this achievement test situation. It does result in a much more controllable situation. For example, Mr. Langmuir has mentioned what a horrible mess we would get into if we followed Buros' advice. Well, I grant you the

job of the test-constructor would become much more difficult.

I am very much of the opinion that in much of the testing work being done today with these emasculated instruments we are just enjoying a vast illusion about grade progression and differences in achievements, and so on. It seems to me that it is very obvious that, if you assume something, you can't use that measuring instrument to determine whether or not that assumption is valid.

MR. LANGMUIR: I would like to make a constructive suggestion to resolve this problem. I have already suggested that we can't get along with one score, but ought to have subscripts, superscripts, and other scores. Now, I suggest that we can't get along with one test. We should always produce our tests in pairs. Our first test will be the maximum-differentiation test, in which the items are selected to discriminate among individuals as best we can. That will be the valid test by our current notions. Along with that we will have the nonvalid one of a matched pair of tests. This will be a test of all those items which definitely do not differentiate among the grade levels. If inspection of the nondifferentiating test reveals to teachers many items of real curricular importance, I feel sure that the teachers will remedy the situation. They will teach the subject and thus make the items valid.

CHAIRMAN FLANAGAN: I think it is about time we threw the discussion open to the floor. But before we do that, I would like to suggest a way of thinking about validity which may clarify some of the issues involved.

There are three senses, I think, in which validity is frequently used. One is, how well does the test represent the real situation? By that I mean the real life situation that we are trying to predict. Second, how well does the test represent the universe from which the sample of items was drawn? This relates to Mr. Buros' complete specifications as to just what you are trying to measure with a test. What is the workmanship? How good a job has been done in measuring whatever the detailed outline specified the test was to measure? The third sense in which validity is used is: How well can we predict something from the test? In this last sense, that of predictive validity, the test might not have any apparent similarity to the behavior that we are trying to predict at all, but through statistical procedures find that it is associated with it.

The discussion will proceed with remarks from the floor.

DR. WALTER DUROST: As the person who was more or less directly responsible for the work done on the Metropolitan Achievement Test and, before that, on the Stanford Revision, a number of the sugges-

tions that Mr. Buros has made are naturally of great interest to me. Only very recently I have crossed the fence into those green academic pastures where I can afford to be unrealistic in my criticisms in connection with testing problems. Until this past September I had the problem of making tests that would be used in the public schools in this country.

I wonder if Mr. Buros realizes—I am sure he must—that if we were to follow his plan we also would have to throw out all of the methods of interpreting test scores which we now use. If we had no gains in score from grade to grade, obviously we could have no grade equivalents.

I wouldn't quarrel with him as to the desirability of abandoning grade equivalents in many cases, certainly in the case of the history test which he mentioned. I would be very happy indeed to see them abandoned, although the test publisher has to more or less go along with his public in a situation like that and give them grade equivalents when they insist on having grade equivalents. But it seems to me he does miss the point when it comes to subjects such as arithmetic or spelling, where there is a continuity of instruction from one grade to another. There, it seems to me, it is entirely valid to investigate the increase in the percentage of individuals answering the questions correctly from one grade to another and to use that as a criterion for determining whether or not an item should be used.

There are many other points which I would like to comment upon, but I notice a number of other people in the audience who have some questions to raise, so I shall only point out one other fact. If we were to abandon the procedures which we currently use in making up achievement tests, we would have tests which, according to the criteria we now use, would be very unreliable. We have to deal with the practical situation of having very limited classroom time for testing. If most of our items did not differentiate between grade groups, being in effect nonfunctioning items, the range of scores on our tests would be seriously restricted, reliabilities would be very seriously lowered, and I wonder what publisher would venture to issue a test to use in the schools under such circumstances?

DR. HENRY DYER: Mr. Buros said he doesn't like the assumption that you construct a test on, differentiating from grade to grade. You have to build a test on some assumption, don't you?

DR. BUROS: Yes, sir.

DR. DYER: I want him to make explicit the assumption he tries to build on.

DR. BUROS: I can say it in a sentence or two. The assumption is

that we have educational objectives. We must try to define those educational objectives in the terms of sorts of behaviors which we wish to bring about in children, and then set up our test to attempt to construct measuring instruments which will give us evidence about growth or lack of growth toward those particular objectives.

I should like to reply to Mr. Durost, but I know the time is going on.

MR. EUGENE CARSTATER: It seems to me that we need to go back to the conference of two years ago and bring some of the ideas that were expressed there to bear upon the discussion that has been going on here. If I remember correctly, one of the ideas that was stressed two years ago was that a test is not valid *of* something but valid *for* something. In other words, it is not what you put into the test, but what you get out of it that counts.

What does a score on a test mean? That is taking up Langmuir's point, I suppose. Does the test tell you the things which you need in order to make use of it? It seems to me that, when we have been talking here about validity in these general terms, we have been talking very much in terms of what various uses can be made of some—I don't know what—standardized test, regardless of whether it was designed for those purposes or not.

One aside on that. Some time ago one of the Navy officers in a postgraduate course in personnel administration wrote a term paper in which he said that the G.C.T. was the most valid test of the Navy Basic Test Battery!

DR. J. WAYNE WRIGHTSTONE: I have listened with interest to the discussions. I have two points that I feel are basic to a clear and logical discussion of some of the disagreements expressed here. I am raising these points from a practical, classroom point of view, with which I must deal. First, we have talked about achievement tests, and I have assumed from the discussion that the main emphasis has been upon standardized achievement tests; yet, in a survey of testing practices among modern teachers, investigators find that at least eighty per cent of classroom achievement testing, regardless of its validity or reliability, is done by means of informal, or teacher-made, achievement tests which are adapted to their particular curriculum and to such objectives as they feel they can measure. Criteria for construction of informal achievement tests is a vital issue that has either been ignored or not brought clearly to the surface in this discussion. This issue needs clarification.

The second point is one that has been implied in one or two other remarks: In our analysis of so-called achievement tests, there

needs to be recognition of the desirability, for tests of reading, arithmetic, and certain of the language skills, of differentiation and of progressive difficulty of subject matter with increasing pupil maturity. We have found that certain of the standardized tests, with items differentiating for age and ability, are valid for the public-school curriculum. On the other hand, in our selection of standardized achievement tests, we have found certain tests of social studies, science, and the like, in which the items are of questionable statistical validity (and we have not ordered these tests because they do not fit our curricular validity in New York City). I know we cannot explore these points more fully at this time, but I feel that unless these questions are faced very frankly, and unless we go back to Dr. Rulon's proposal and define more clearly what is a habit, what is a so-called ability, or what is any of the other psychological concepts, we shall increase our confusion, rather than decrease it, on the problem of validity in relation to item analysis and in relation to other statistical techniques which may be employed.

PANEL II

Defining the Nature of the Verbal Factor as Used in Aptitude Test Batteries

Why a Verbal Factor?

by IRVING LORGE

In the development of devices for the appraisal of intelligence, a large variety of different test forms have been used. Among the many that may be listed are: analogies, logical absurdities, classification of words into categories, sentence completions, synonyms or antonyms, word meaning, reading comprehension, naming objects, giving rhymes, saying as many words as you can, and inferring the properties of a set of words.

These test forms have at least one common attribute: each uses words in one way or another. Sometimes they deal with syllables or even single letters, although, as is to be expected, more frequently the test items are concerned with whole words, sentences (complete or incomplete), paragraphs, or larger units.

Perhaps there is some sense in putting together by appropriate means the results of an individual's operations in getting the common principle for a group of related words, in the garrulousness or fluency in saying unrelated words, and in giving rhymes for a set of syllables. The score, then, would be based on the community of content.

But is community of content what ought to be measured? Is there not some need to recognize that intelligence reflects the acquisition, development, and communication of ideas, that it deals with the processes of getting and holding ideas, of reasoning and utilizing them, and perhaps of inferring or imagining them. The emphasis, in my opinion, should be more upon process than on content. When the factor analysts can get from the same matrix not only factors identified with number, word, and space, but also with deduction, induction, perception, and reasoning, just to name a few, the problem seems to be how they do it. Heisenberg's principle would suggest that there is at least an element of uncertainty and indeterminacy in identifying components for content at the same time as for process.

At the risk of completely confusing the issue, it must be remembered that words may be used in various ways: in general some test interpreters distinguish between words used abstractly and those used concretely. Most of you can recognize that some presumptively nonverbal tests—such as the Goldstein, Vigotsky, Gelb, and Weigl—are concerned with behavior which may be classified as concrete or abstract. It is even alleged that as chil-

dren develop they move from concrete to abstract behavior; and, it may be added, it is also alleged that in the deterioration of some adults, the performances become less abstract and more concrete. Sometimes, of course, the concreteness is so uniquely in the individual that we say communication is, to a degree, suppressed. Therefore, it must be recognized that some of the processes involved in verbal test content may be used in other kinds of content. The correlations between so-called verbal and nonverbal tests vary. In a study by Leon Brody the intercorrelations among verbal abstract and concrete tests and nonverbal abstract and concrete tests ranged from .69 to .84. The correlation of the four subtests with an intelligence quotient derived from a group test was not that high; for the verbal concrete, it was .52; for the verbal abstract, .54; for the nonverbal concrete, .40; and for the nonverbal abstract, .45. Such intercorrelations imply that when there is a community of process, the relation among test devices will be relatively high, and as two test devices appraise intellectual processes increasingly different from each other, the correlation decreases, even within the same content area.

It seems, therefore, that modern intelligence tests should stress the processes underlying intellectual performances. While it is easier to make appropriate items in verbal materials, similar processes can be appraised in the content areas of number or of space. To the degree that similar processes with similar weights are tested in various content fields, the necessity for separate itemization of verbal, number, and space factors will become less attractive.

A departure from the heavy overload of verbal content in intelligence tests is needed to get a better understanding of the average behavior of individuals not only with ideas, but also, using E. L. Thorndike's concept, of manipulating things and dealing with people.

Vocabulary as an Arbitrary Definer of the Verbal Relations Factor

by EDWARD E. CURETON

It seems fairly clear that the discussion title of our panel: "Defining the nature of the verbal factor as used in aptitude test batteries," poses an insoluble problem, since we already know with reasonable certainty that there are several distinct if not independent verbal factors in the tests which are commonly found in aptitude batteries. Thurstone's first major study revealed two, and several more have since been reported.

I take it, however, that we are concerned primarily with what has been termed the "verbal relations" factor. Most of the tests which have high loadings on other verbal factors are tests which do not commonly appear in aptitude batteries. I wish to suggest that it may prove serviceable to define this factor simply as *vocabulary*. This suggestion is not new, and it is not originally mine. It was advanced several years ago by Dr. John B. Carroll. Whether or not it was original with him, I do not know.

This suggestion appears to be reasonably consistent with the facts. In factor studies involving large numbers of tests, there is usually some latitude for personal bias in making rotations, even when we allow these to be oblique and try to satisfy the criterion of simple structure. If we keep them orthogonal and allow some departure from simple structure in order to avoid second-order factors, the latitude is even greater. It should, therefore, be possible in most cases to rotate in such a manner that the "verbal relations" factor absorbs all the common-factor variance of the vocabulary test, without doing much violence to the rest of the data. This provides us with a factor for which there is an arbitrarily defined "pure test."

The more usual alternative is to rotate in such a manner that the "verbal relations" factor has more or less equal loadings on vocabulary and such other tests as mixed sentences, verbal analogies, proverb matching, paragraph comprehension, and the like. When this is done, the vocabulary test has just about as high loadings on the reasoning factor as do these other tests. If we rotate first in the manner suggested previously, we should then be able to obtain reasonably clear answers to two fairly important questions:

1. Do the loadings of the other verbal tests on the general reason-

ing factor increase as a result of this procedure?

2. Do these other tests then proceed to generate a "verbal reasoning" factor which is distinct from the factor commonly associated with such tests as arithmetic reasoning and figure classification?

These questions seem clearly worth answering; but, so far as I know, no studies of the type just suggested have yet been made. I am sorry that I have not been able to make one or two myself in preparation for this conference. To make them it will only be necessary to re-rotate existing factorial matrices which are to be found in the literature. Meanwhile, I should like to mention a few minor studies involving verbal tests only.

One of these involves only four tests: a reasoning test which is mainly but not entirely verbal, modeled after the one devised many years ago by Burt; a mixed-sentences test; a five-choice vocabulary test of the usual type; and a synonym-antonym test containing also class-member pairs (e.g., animal-cow) and pairs not related in any of these three ways. This last *appears* to be essentially another word-meaning test, and has been given that name in the study. In a larger battery in which these were the only strictly verbal tests, all four had more or less equal loadings on the "verbal relations" factor, and all of them except mixed sentences had high loadings on a not-too-clearly defined reasoning factor. The vocabulary test and the synonym-antonym test generated a small doublet in addition. When considered separately, however, the tetrads generated by these tests were all close to zero, namely –.011, –.035, and –.046.

There is only a hint of a group factor in either the two vocabulary tests or the two verbal reasoning tests. The addition of an alphabetizing test, which also had a high loading on the "verbal relations" factor in the larger battery, likewise gave rise to low tetrads, the highest of the twelve being .098 and the average, –.008. The standard errors of these tetrads, which are based on a sample of 400, are of the order of .01. Hence, the larger ones are statistically significant, but none of them can be called high. The distribution of the tetrads suggests a vocabulary group factor and perhaps one or two others, but the evidence is not at all clear.

A second study included the same two vocabulary tests, along with a verbal analogies test, a paragraph reading test, and a proverb-matching test. In this study the experimental design was set up specifically to permit a complete tetrad and triad analysis. The sample was larger—841—and the result was unequivocal. There was a clear group factor in the two vocabulary tests, but none in the three verbal reasoning tests. Since the correla-

tion between the two vocabulary tests, when corrected for attenuation, was .98, it was apparent that this group factor was an overlapping specific factor rather than an independent trait.

A third study consisted of a reanalysis of Davis's data on the Cooperative Reading Comprehension Test. In this analysis one factor was forced to coincide exactly with the vocabulary test. This factor, so defined, absorbed three-fourths of the nonchance variance. A "verbal reasoning" factor appeared also, along with two others of equal magnitude which were apparently due to nothing more than similarities of item-structure in certain of the tests. Each of these three other factors absorbed about 1/12 of the nonchance variance.

These studies at least suggest that verbal tests of the types used in aptitude batteries, when taken alone, generate only one major factor, and that the vocabulary test is at least as good a measure of this factor as any other. Questions concerning a "verbal reasoning" factor, and its possible relation to some more general reasoning factor, have been raised but not answered.

If we are interested merely in the construction of aptitude test batteries which will measure the largest number of factors with the smallest number of tests, rather than in theoretical problems concerning the nature of the human mind, two methodological suggestions seem to be worth considering. The first is that the Spearman-type techniques still have important values whenever we can deal with small enough groups of tests to use them. The second is that certain problems may profitably be studied by what we may term the "method of biased rotations." In the case of the "verbal relations" factor, it may prove profitable to study residual factors after forcing this factor to coincide exactly with vocabulary.

Delimiting the Verbal Factor

by FREDERICK B. DAVIS

Everyone recognizes that high ability in the verbal factor denotes facility with words and that low ability in this factor denotes the reverse, but not everyone has precisely the same conception of what facility with words means. To some, it connotes ability to remember word meanings and to make increasingly fine distinctions among them. In her diary, Dorothy Wordsworth has recounted an amusing conversation between Samuel Coleridge and her brother regarding the best adjective to use for describing a waterfall called the Linn. Coleridge hesitated among *sublime, majestic,* and *grand.* He was making fine distinctions among the precise connotations of those three adjectives, any one of which might well have completely satisfied a less sophisticated literary craftsman, or, to put it in the terminology of measurement which we are using this morning, men with lower scores in the verbal factor.

A measurement of memory for word meanings may be accomplished by means of test items like the following:

inviolable

A unapproachable
B undefiled
C virtuous
D not violent
E sacred

The degree of precision with which distinctions among words is measured can be increased simply by incorporating decoys in the items that are closer and closer to the exact meaning of the correct answer. Unfortunately, many tests of this kind have obtained greater and greater difficulty in their component items by employing esoteric words with decoys of little plausibility rather than by employing useful words with ingenious and attractive decoys. Examples of esoteric words that have been used are "syzygy" (which is the last word in the vocabulary test of the *California Test of Mental Maturity*), and "umbel." These are typical of the highly specialized words that have sometimes been included in general vocabulary tests used to measure the verbal factor.

Some people have thought of the verbal factor as ability to determine the precise meaning of a word from its context. This has led them to avoid verbal items of the kind cited above and to test the meanings of words as they appear in sentences or even in paragraphs or stories. Ordinarily the context provided by short phrases or sentences is mere window dressing that serves no useful purpose, but I have presented

evidence that a mental skill independent of memory for word meanings does indeed seem to exist when genuine dependence on context is actually required for obtaining the meaning of an unknown word, or for obtaining the meaning of a word like l-e-a-d, which may be pronounced "lĕd" or "lēd" depending on the meaning indicated by the context.

A third group of people have conceived of the verbal factor chiefly as the ability to manipulate words in combination and to reason with them. Measurement of these abilities has often been accomplished by means of items like the following:

Assimilation is to vitality as exertion is to

A exercise.
B fatigue.
C satisfaction.
D energy.
E decay.

Needless to say, there are other concepts of the verbal factor, as has been mentioned. Dr. Thurstone, for example, has noted the existence of a verbal factor which he has described as word fluency.

I would like to see the verbal factor defined rather narrowly as an index of both range and precision of word meanings and to suggest that it be measured by means of cleverly contrived multiple-choice items in the format of the first item quoted above. The verbal factor would then be divorced in large measure from the reasoning abilities involved in items of the word-analogy type or of the opposites type. These reasoning abilities are probably important enough to warrant assigning separate names to them. They may be so nearly identical with the reasoning-in-reading factor identified in two separate batteries of reading tests that they can be conveniently included in that factor's variance.

It is likely that measures of both the verbal factor and of the reasoning-in-reading elements, as I have defined them, will prove to have substantial social utility. The contextual-meaning skill may also prove to be worthy of separate measurement. The extent to which these separate measures correlate to markedly different degrees with various criteria of real consequence will, however, determine whether they are worthy of separate measurement.

Verbal Factors

by ALEXANDER G. WESMAN

I am afraid that I do not know what *the* Verbal Factor is. When I worked with the Graduate Record Examination, we leased three verbal tests from the College Entrance Examination Board—a vocabulary-opposites test, a double-definitions test, and a paragraph-meaning test—and called the total score from these three our Verbal Factor. Thurstone has proposed V (Verbal Relations) as a verbal factor, but he also recognizes his W (or Word Fluency) factor as being verbal in nature. In the Differential Aptitude Tests we have a Verbal Reasoning Test, but we have a Language Usage Test, composed of Spelling and Sentences, which also contributes to measurement of verbal ability. Further, if the factor analysis research on Mechanical Comprehension in the Air Corps, as reported by Guilford, can be generalized to our Mechanical Reasoning Test, then the MR test, too, has some saturation in the Verbal Factor as it is defined there.

Perhaps the foregoing is evidence only of the impurity or factorial complexity of the several tests. An alternative explanation which seems to warrant consideration is that perhaps *the* Verbal Factor is not yet sufficiently defined or understood. It may well be that the course of wisdom is to recognize the possibility that there is not one Verbal Factor, but several Verbal Factors. There is increasing evidence that memory, which was regarded as a unitary factor for some time, must be recognized as several factors. It seems equally possible that continued research will demonstrate that there are several Verbal Factors as well. There is strong suspicion in the minds of many of us that in discussing the Verbal Factor we are reifying a mathematical concept, that in trying to make mathematically derived factors psychologically meaningful we have sometimes acted as though our labels gave them psychological existence and character. I, for one, am not at all certain that the Verbal Factor as derived from different sets of tests and different populations is psychologically invariant.

Neither am I certain that the factors which result from analysis of test results alone are the best measures of the ability or abilities with which we may be concerned in aptitude studies. I have in mind some results reported in the same Air Corps volume to which I previous-

ly referred. In that volume reasons were discussed as to why the Verbal Factor was necessary for pilots, bombardiers, and navigators. We are then informed that their vocabulary test is "practically a pure measure of the Verbal Factor" and that "in a battery where measurement of the verbal factor is required, a vocabulary test is strongly to be recommended."

As we continue, however, we learn of a startling development:

> Owing to its limited predictive value, the Vocabulary Test was dropped from the classification battery in the summer of 1942. It was replaced by the Technical Vocabulary Test, which is a test of specific technical information pertaining to piloting, navigation, and bombardiering. This test, although related to vocabulary, and loaded with the Verbal Factor (0.41 for the pilot score), possessed a validity of .21 for pilots. Part of this validity is derived from the test's loading with the mechanical experience factor (0.39). The remaining part is accounted for by its loading with pilot interest (0.34).

It appears then that when this pure measure of the Verbal Factor was exposed to validation against a criterion, it lost much of its luster. Yet it is the very process of validation against an outside criterion which must be the final yardstick in determining the value of the Verbal Factor or any other factor in aptitude testing. We have heard much in the last few years about factorial validity. This may be a useful concept for theoretical research purposes; it is not a substitute for empirical validation. Tests constructed in the light of factorial research must be subjected to empirical validation just as much as tests built according to different principles.

It is my belief, then, that the external criterion must determine the nature of the Verbal Factor as used in aptitude test batteries. If the criterion is one which calls for ability to read and understand technical writing, a test of that ability may be the best measure of Verbal Factor for that criterion. If a knowledge of words and their meaning is important, a general vocabulary or technical vocabulary test may be best. If the ability to reason with verbal constructs is required by the criterion, a verbal reasoning test may provide the greatest situational validity. All of these tests may reasonably be considered as measures of the Verbal Factors.

The problem which faces the aptitude-test constructor is that of preparing a test or group of tests which will predict best in most applications. He may approach his task through factor analysis, or job analysis, or both. He should be swayed by efficiency, face validity, and similar practical considerations; but these are of secondary importance. The prime consideration is still that of building a test which will provide most empirical validity in the greatest number of situations—which will best predict the largest number of criteria. For it is the criterion which, in the last analysis, will define the worth of the factor or the test.

The Relationship Between Verbal Factor Scores and Other Variables

by WILLIAM W. TURNBULL

Preceding speakers have pointed out that studies of "the verbal factor" have shown it to be much more complex than had been supposed originally, with an undetermined number of factors (at least 4 or 5) appearing where only one was originally known.

The immediate problem for attack is, presumably, the surer identification of these factors involved in the use of words. Work of this kind will provide an important part of the answer to the problem implied by the title of this discussion: "Defining the nature of the verbal factor. . . ."

It will be my purpose to present briefly four approaches other than the refinement of the factors that can be isolated through studies of the intercorrelation of written verbal tests—four approaches which I believe will be required in reaching a more complete definition of the nature of the "verbal factor"—or, should we rather say, the "word factors." These approaches will constitute attempts to describe the word factors in relation to external variables as well as in relation to each other.

The first approach is the extension of studies to include media of communication other than writing. Passing mention of this need has been made by Taylor[1] in the discussion of his study, and some work has been done by British investigators in tracing fluency through speech as well as in written work. Studies are needed to determine whether or not the factors found in speaking and in the comprehension of speech are similar to or the same as those found for tests based on reading and writing.

A second approach is the study of the genetic development of the word factors. Studies relating age and the organization of mental abilities have, of course, been made by Thurstone and by others. As new abilities or characteristics are brought to light, however, there is a need for repetition of such studies; and we now require investigations of the word factors as they are exhibited by very young children. At what age can we differentiate these factors, and how do their interrelationships change with age?

[1] Taylor, C. W., A factorial study of fluency in writing. *Psychometrika*, 1947, 12, 239-262.

A third approach is the study of the relation between cultural influences and the development of the word factors. We should discover whether or not the same factors are found in different cultures or sub-cultures. Does factor W operate in verbal tests prepared in oriental languages? ("W", you will recall, involves the ability to produce words which have some formal characteristic in common, such as the same first letter. Is there an analogous ability in languages based on independent pictorial units rather than on an alphabet?) Then, within a society in which a given set of word factors has been identified, we need to know whether or not some sub-cultures are more propitious than others for the growth of word abilities, in general and differentially. It is fairly well conceded that there is a large general effect of experiential background on word skills. Is this effect more pronounced for some word abilities than for others? Are some environments favorable to the development of one of the word abilities and unfavorable to the development of others? A study of systematic variation between word abilities in relation to experiential background would provide an effective supplement to the evidence gained from factor analysis as to the independence of the word factors.

A fourth type of investigation will be that in which an attempt is made to relate word factors to personal characteristics other than mental abilities as usually conceived. Here the work of Stephenson should be mentioned. In an article published some 15 years ago Stephenson[2] indicated positive results in an attempt to relate fluency scores and types of mental disturbance. The manic phase of manic-depressive psychosis had particularly strong relations with fluency, as might be hypothesized. Cattell[3] felt even earlier that written fluency tests were strongly related to a characteristic which he named "surgency"—a temperament trait in terms of which, of course, normal as well as disturbed subjects can be described. In the light of our growing knowledge of the dimensions of the word-factor field and the refinements in testing instruments made possible by the recent factor studies, this earlier work should perhaps be re-done in an effort to relate temperament traits to the more specifically delineated word factors.

When these approaches to the definition of the verbal factor (or word factors) have been taken, we will, I believe, have a good picture of its dimensions and of its place in the psychological economy.

---

[2] Stephenson, W. and others. "Spearman factors and psychiatry." *British Journal of Medical Psychology*. 1934, 14, 101-135.

[3] Cattell, Raymond B. Temperament Tests. II. Tests. *British Journal of Psychology*. 1933-34, 24, 20-49.

The Verbal Factor from the Consumer Point of View

by HERSCHEL T. MANUEL

Taking advantage of the wide latitude suggested by the topic proposed for the panel, I shall enter the discussion from the standpoint of the consumer rather than the test-maker or the factor-analyst. What kind of verbal factor in an aptitude test battery would be most useful to the consumer? By consumer I mean the teacher who wishes to use a battery of tests to discover the abilities of the pupils enrolled in his class, the counselor who wishes a profile of aptitudes for vocational guidance, the personnel officer who wishes to discover the possibilities of employees in relation to job placement and training, and the like. The research worker is omitted from this list on the assumption that he will choose or construct tests for his own special purposes. I am using the term *factor* in its general sense and not necessarily as the lowest common factor or element.

(1) The first point to be made is that the total number of factors to have a place in the profile provided by the whole battery should be large enough to include the aptitudes that are most significant in relation to the use to be made of the tests and yet held to the minimum necessary for the purpose to be served. Human nature is complex and oversimplification is an ever present danger. The urge to learn all about an individual in a few short tests is hard to withstand. On the other hand, time, expense, and the sheer difficulty of interpreting a large number of variables call for the minimum list that will do the job.

(2) Tests should be chosen in such a way that the score on the verbal components of the battery, or some simple derivative of this score, will represent the verbal factor without further analysis. By the same token other factors in the profile should be based upon materials which will yield scores not too greatly distorted by uninvited verbal elements.

(3) The verbal factor should be expressed in language easily understood and accepted by the consumer as reasonable. A recent explanation from a bulletin of the College Entrance Examination Board is offered without evaluation as an illustration: "The verbal

section is designed to measure the candidate's understanding of words, his skill in dealing with word and thought relationships, and his ability to read with understanding and discrimination." (Bulletin of Information for January Series, p. 8.)

(4) The materials used should have face validity in the light of the definition. That is, the consumer should be able to see that the test materials conform to the description.

(5) The verbal factor or factors included in the profile should be of maximum value in relation to the purposes to be served by the battery. It may be, therefore, that batteries for different purposes will have different verbal factors. Some purposes may need an emphasis on particular types of verbal material or a more analytical approach than others.

(6) To assure maximum validity the verbal section of the test should include varied components of relatively short length rather than longer sections of less varied materials. This is merely an application of a general principle emphasized by John Dailey in his recent APA paper.

(7) In the interest of comparability a standard verbal factor may be defined and then used by different authors working in the same field. This definition should probably be based upon the types of materials to be included and the test form in which the materials are to be presented. Obviously, this proposed uniformity should not be allowed to crystallize the *status quo* and to discourage further experimentation in a field in which our knowledge is still so meager.

DISCUSSION

Participants:

IRVING LORGE, EDWARD CURETON, ALEXANDER WESMAN, JOHN CARROLL, HERBERT CONRAD, FREDERICK DAVIS, ANDREW R. BAGGALEY.

IRVING LORGE: I am particularly pleased that Dr. Manuel made the point about the consumer approach to the factor defined as the verbal factor. There are two points that ought to be made about the verbal factor. First is the point that the use which people can make of the verbal factor depends greatly upon how it is appraised. I use the word "appraised" in preference to the word "measured" because, as I stated earlier, measurement cannot distinguish *simultaneously* between process and content.

The second point is that the nature of the scores obtained from various tests of verbal factors sometimes adds more confusion. For instance, there is a significant difference between the abstract and the concrete use of words and the abstract and concrete handling of objects. In a recent study, Dr. Herman Feifel demonstrated several levels in the understanding of words. As Chambers originally demonstrated in the early 1900's, the earliest definitions given by children are in terms of use. As children grow older, words tend to be interpreted in terms of explanations or of demonstrations; and then farther up the age scale definitions are given in terms of genus-species or of synonyms. The result is that in the kind of vocabulary score obtained from tests credits are indiscriminately added for answers based on use, or on synonym, or on explanation, or on demonstration, as if each variety measured an identical concept of vocabulary. The vocabulary score is therefore a vague confusion. It is an undifferentiated mixture of the subject's culture, his genetic development, or even of his deterioration. We must remember that the nature of the score that goes into factor analysis is not a very pure or precise number. Yacorzynski emphasized, in his critique of the Babcock level efficiency theory of measuring intelligence, that the vocabulary score taken as evidence of innate intelligence may be made up of different behaviors at different age levels.

EDWARD E. CURETON: I should like to make a few remarks con-

cerning Dr. Wesman's presentation. It seems to me, if I understood him correctly, that he interpreted data from Guilford in such a manner as to indicate that his own company's differential aptitude test, or at least his own particular verbal portion thereof, was no good. I would take issue with that. It seems to me that a single general measure of the verbal factor or of any other useful factor may be very valuable.

It is true, of course, that if we are interested in the prediction of one single external criterion, the thing to do is to make up tests directly related to that criterion. This will of course present lots of work for the test-building profession for years to come. On the other hand, I think it is possible still that an aptitude battery can be built which will be valid over a rather wide range of different criteria. That at least is the use, if any, of factor analysis in the construction of aptitude batteries.

We try to find, fundamentally, the smallest and most suitable number of tests to measure the widest range of aptitude combinations. To do that ideally, we should find the separate or independent (or as nearly as possible independent) factors, and measure each of them with one test. That is why, from the practical standpoint, I object to any form of factor analysis which defines every factor in terms of a combination of a lot of different tests, and why I feel that we can quite usefully define factors in terms of single tests according to the criterion that, when we are all done, the single test shall have low intercorrelations, and each one of them shall correlate with some one or more important criterion measures.

In Dr. Wesman's presentation he suggested technical vocabulary tests and various others which seemed to me to be essentially information tests. I think that, in the case of the verbal field, we should distinguish between an information test and a general aptitude test in this way: An information test must of course involve the medium of communication as well as the information; the verbal test might conceivably measure command of the medium of communication without attempting to measure specifically its use in learning or returning information in any one particular area.

ALEXANDER G. WESMAN: I certainly have no quarrel with Dr. Cureton as to the desirability for having your aptitude test measure in as many areas as possible. I have even less quarrel with him, naturally, when he says the particular test of which I am a coauthor is pretty good.

I think what I want most to emphasize (rather than quarrel with Dr. Cureton, since I would agree with most of what he says) are two points: One is the danger that I feel, in reading some of our current

literature, of accepting factorial validity in place of empirical or situational validity: the notion that once you have a test which is a so-called pure measure of a certain identified factor that you need go no further. I am not ascribing this notion to Dr. Cureton; I think he would be fully in accord with my views. It is rather a general caution.

The second notion, and one on which Dr. Cureton may want to comment further, is the idea of information versus the essence which a factor represents. I think if you accept my definition of aptitude, namely that it is anything which will predict success in a given vocation or educational realm, that frequently your best measure of aptitude may turn out to be an information or achievement test. Certainly I see achievement tests as being aptitude tests in much of their use. If we take something as common in the industrial field as the study of stenography—and this applies as well to the academic field in commerce courses—I have yet to see any test, however derived, surpass a spelling test in terms of prediction of success at learning stenography.

CHAIRMAN FLANAGAN: We don't seem to be able to stir up any very great controversies among the panel members. We will throw the discussion open to the floor and see if some people have views divergent from those expressed.

JOHN B. CARROLL (Adjutant General's Office): Since there has been mention of my suggestion that the verbal factor can be defined as essentially a vocabulary factor, I thought I might take a little time to amplify the suggestion.

It came to my attention that many of the tests which Thurstone was using to measure this verbal relations factor, as he called it, of course also had some loadings on the reasoning factors. Then I began to analyze the situation. I must say here that I am interested in *factorial* validity. I am interested in analyzing the fundamental dimensions; the matter of empirical validity can come later.

I analyzed the situation in this way: Most of our vocabulary tests as presently constructed involve many words, as Dr. Davis pointed out, which are of the nature of "syzygy" and comparable words at the very high level of vocabulary. Now, how do we develop our knowledge of those words? Well, it is usually developed in our reading by learning the meanings of words through their contexts, and I suspect therefore, that in the learning of those words of high vocabulary level, reasoning will naturally be involved. So, in defining the verbal factor apart from its reasoning aspects, I was trying to find measures of *knowledge of linguistic units*, let's say, not necessarily knowledge of words alone. The verbal factor may have to do with lin-

guistic phrases, or with the use of certain grammatical elements of our language. This is why, of course, grammar tests usually come out on this verbal factor. But my main evidence for the suggestion comes in the apparent factorial purity of a test which I called the Phrase Completion test, which involves simply scoring a set of incomplete phrases in terms of the community of response to those phrases. The phrases are such things as "Hounds and ——," or "What's the——?". Those phrases presented to a group of people result in certain completions which are then scored for community of response. In at least three factorial studies, that test has come out with a rather high factorial validity for the verbal factor and a pure factorial structure. The responses involved in this test are obviously linguistic units of a low level of "difficulty."

If you are interested in further elaboration of this point of view, I would refer you to an article on "The analysis of verbal behavior," which I published in the *Psychological Review* in 1944.[1]

I wanted to speak about a few other points raised here. First, as to this problem of the technical vocabulary. Here again if you consider the verbal factor as the "centroid" of *knowledge of linguistic units*, you will see that in the use of English we have such a "centroid" in the common words and phrases that are used in everyday life and which are normally used in even technical language. For example, the words "hypothesis" or "condition" are words of rather general application and we are concerned with the centroid which contains such words.

Now, when we get into special vocabularies such as the language of biology when we talk about metabolism or protoplasm, and so on, we are obviously going to have sub-centroids, as it were, or independent centroids, and I believe there are some studies which show that you get linearly independent factors or centroids in such vocabulary tests.

Then on the problem of cross-cultural investigations, I can only point out that there are some factorial studies available which concern tests in foreign languages. There is, for example, a study by Meili of Switzerland, with whom I was able to talk this summer at the International Congress; he gave me an extensive correlation matrix, which he had recently developed, which clearly showed some W factors quite as strongly in the use of languages which are fairly similar to ours, as in the test batteries using English.[2]

HERBERT S. CONRAD (U. S. Office of Education): The vocabulary

[1] *Psychological Review*, 1944, 51, 102-119.

[2] Meili, R. "L'Analyse de L'Intelligence." *Arch. de Psychol.*, 1946, 31, No. 121.

test is a test of high repute. I think it has perhaps gained even more repute in the discussion this morning; and yet the College Entrance Examination Board makes no use at all of the vocabulary test.

There are at least two objections to the vocabulary test that seem to warrant attention. First, its coachability, the fact that vocabulary is teachable. It is quite possible for an investigator to spend a lot of time and money developing a valid vocabulary test which would become increasingly invalid as the years go on, due to the fact that teachers and other persons undertake to coach the students who are going to take the test, and thereby reduce its validity.

The other point relates to consumer acceptance. There has been a great deal of discussion for many years about using information not merely for testing how much a person knows, but also to determine what use he can make of it. In the case of vocabulary, the application of that principle is obvious: teachers want to know whether students can use the words. Thus, it is not enough for a student to know the individual words in a reading passage; he must be able to understand them in context, follow the author's meaning, and so on. If we go too far and thrust the vocabulary test down teachers' throats with the statement "We have proved it is valid," we will be going beyond the tolerance of our customers; and, while I don't think that we should toady to the customers, nevertheless I think that one has to give some recognition to their opinions and lead them along gradually.

DR. LORGE: I have been listening to this discussion of vocabulary for some time. I want to confess that I have been responsible for counting words for a long time. Dr. Thorndike and I have counted over 25 million running words so that we were able to list the 30,000 most frequently occurring words. If any student could be coached on all of the meanings of those 30,000 most frequently used words, he would have a vocabulary score that would mean something.

One has to recognize that vocabulary is a measure of something. What it is a measure of, is not very clear. We know that it is a better measure of itself than are any of the present measures of democracy, intolerance, or prejudice. In discussing vocabulary, at least, we think we know what we are talking about.

In the Institute of Psychological Research, I have computed the intercorrelations among different kinds of vocabulary tests: first, there is the usual five-choice vocabulary test. In such an ordinary vocabulary test, we are not testing the meaning of one hundred words, but of six hundred words, because the respondent must know the meaning of the key words, and at

least some of the meanings of each of the choices, right or wrong. We, as yet, do not realize how much we are testing in vocabulary tests.

The second kind of vocabulary test has the usual stem word and several options, including one or more of the different meanings of the stem, as *game* means both *food* and *play*. The respondent indicates all those options which he recognizes as acceptable meanings of the stem.

The third type of vocabulary test item is one in which the subject indicates whether a word is used correctly. In this test type the subtle distinctions that Dr. Davis discussed are appraised. The fourth test type is a triple-definition form, in which definitions must be matched with three stem-words.

From each of these four types of vocabulary tests we get a score. The intercorrelation among the four scores for elementary school children is so high that there is no evidence for claiming that one is a better measure of vocabulary than any of the other three. The factorial purity is not only apparent, it is demonstrated. The problem becomes which *form* will you use? In general, the one that is easiest to use. You have a choice of asking for discrimination among word meanings on a subtle level which is the equivalent of asking for the meaning of "syzygy" on the frequency level. If you are asking a person to distinguish, let us say, between *game* meaning *to play* and *game* meaning *lame*, or *run* meaning *to proceed quickly* and *run* as a *stream of water*, you are not really testing anything but relative frequency of word meanings. The fact that they have the same symbolization on paper doesn't make them more or less frequent. The idea that a common word has hard, recondite, or rare meanings was the rationale basic to making the semantic count.

The intercorrelation among these four vocabulary types and reading tests indicates that the verbal factor measured in context does not seem to be fundamentally different from what the four vocabulary tests measure. I, of course, recognize that the flow of words, the ability to regurgitate words without meaning which insane persons sometimes have, is different from the recognition of meaningful relationships with words and among sentences.

DR. WESMAN: I want to take just half a minute to add an illustration to what Dr. Lorge has just said. One of the vocabulary tests with which I have played is one very similar to that which he described. It gave a stem, and then ten options, some of which (anywhere from 0 to 10) were dictionary synonyms of the stem word. The chief reason why I have done nothing with the test since its first and second tryouts is that I have still to explain to myself why I get so very

many negative biserials of the approximate range of .30 and .40. The explanation that I cannot accept is that my criterion score, which was total score on the test, is a poor criterion as such, based, as it is, on the dictionary. It is not a question of false keying. There is apparently some difference in the kinds of understanding of these semantic meanings which is responsible.

DR. DAVIS: I should like, if I may, to try to sharpen the issue. We have had some comments from the floor but I think none from the panel about the value of the vocabulary test. Let's remember that we are *defining* the verbal factor here—not *evaluating* it. I should like to see it defined, as I have indicated, in this fashion as a vocabulary test—that is really what it amounts to—with precision of word meanings being the central element measured. Now, I am not saying, when I ask for that, that the factor is of any social value. It may be and it may not be. I am pretty sure it is for some things. I am sure it isn't for other things. I am sure that people in the academic world greatly overrate its value for most things, but the point at issue is not whether a verbal factor test such as a vocabulary test is a valuable test or is not a valuable test for any specific purpose. The point at issue is whether to define the verbal factor in terms of a vocabulary test. Personally I should like to see this done. Then the other elements in the verbal factor, as we now define it (reasoning, getting meaning from context, and so on), would also be measured separately, named differently so we could make distinctions among them, and evaluated separately. It is my belief that they will prove to have different values for different purposes. If their different values for different purposes are great enough, then the fact that we make the distinctions and measure them separately will turn out to be worthwhile.

DR. CONRAD: I would like to say, with regard to Dr. Lorge's comments, that I have spoken to teachers who have a standing in their profession I guess as high as the panel members here, who are convinced that by teaching their students vocabulary they do a lot of good for their students when it comes to taking the College Board aptitude examination, even though there is no explicit vocabulary test in the College Board examination.

ANDREW R. BAGGALEY (Harvard University): I don't think there is such a sharp separation between the issues of the factorial validity and the validity in practice, because tests which correlate low with each other correlate more highly with the criterion than if they correlated highly with each other, so I think the two issues are joined in this matter.

CHAIRMAN FLANAGAN: We will now adjourn for lunch.

PANEL III

Establishing the Type of Norms Most Useful and Important for the Interpretation of Achievement Test Scores

Value of Norms Based on a New Type of Scale Unit

by E. F. GARDNER

In this brief presentation I propose to point out the need for interval scales for measuring such functions as individual differences and growth in academic achievement, and then to describe some results obtained when newly developed units called K-units have been used.

In spite of the fact that considerable progress has been made in the development of devices for measuring human achievement, it is generally recognized that even the best tests available are crude and ambiguous in comparison with ideal instruments. Scores or norms presented as aids to the interpretation of student performance on tests are frequently inadequate. One of the more important contributors to such inadequacy is the lack of equal units throughout the usable range of the scales. The extent to which it is proper to add, subtract, multiply, divide, and compute ratios with measures obtained from testing is often not made clear. Edward L. Thorndike calls this fundamental defect in an instrument for measuring intellect, "arbitrariness of units."[1] Whenever any scale of amount is used, one's natural tendency is to interpret it as one has learned to interpret common scales of number—to regard every unit in the scale as equal to any other unit and to regard the zero, to quote Thorndike again, as measuring "just barely not any"[2] of the quality in question.

Rigorously, statistics, such as means and standard deviations, should not be used with scores based on an ordinal scale, for such statistics imply a knowledge of something more than relative rank order of data. On the other hand, this nonrigorous, although common, practice seems justified on the grounds that in numerous instances it leads to fruitful results. While prohibiting this procedure would serve no useful purpose, it is proper to point out that means and standard deviations computed on a

[1] Thorndike, Edward L. *The Measurement of Intelligence.* New York: Bureau of Publications, Teachers College, Columbia University, 1927. P. 3.

[2] Thorndike, Edward L. *Theory of Mental and Social Measurement.* New York: Teachers College, Columbia University, 1913. P. 7.

scale based on rank order are in error to the extent that the successive intervals on the scale are actually (not just numerically) unequal in size. For this same reason, growth measure on a scale not having equal units is likely to be in serious error.

A direct approach to the problem of establishing equal units would be to construct each test so that the score would yield an interval scale directly without conversion. However, Flanagan points out that, although theoretically it might be possible to eliminate the problem of units by constructing a test in which the trait under consideration was precisely defined by the content of the specific items included, such a procedure is impracticable.[3] A test practically always consists of only a limited sample of all the items it would be appropriate to use, since the time and the expense of approximating a test which would completely define a trait is prohibitive. It is therefore necessary to look elsewhere for a method of obtaining equal units of measurement.

In science two general methods have been used to establish equality of units: (a) by direct comparison and (b) by definition. Mental traits obviously cannot be superimposed and compared directly, so, like temperature, units of mental measurement must be made equal by definition. There need be no apology attached to such an admission since even in the physical sciences there are few quantities which have units whose equality can be determined directly. The relationship between the units used and the shape of the frequency distribution of a particular trait has been used extensively in the past to define specific trait units.

Following the teachings of Galton and Quetelet, the assumption underlying the search for quantitative units has been that learning is normally distributed and that equal segments of the base of a normal curve mark off equal units of learning. Thorndike has, in many instances, utilized this method of determining units. His CAVD units are one application of this technique. The familiar T-scores of McCall and the more recently developed Normalized or Scaled Scores of Flanagan are both based upon the assumption of a normal distribution of the trait under consideration within a specific group. The T-scores of McCall and the units of CAVD Psychological Examination of Thorndike present illustrations of equal units defined in terms of a single frequency distribution. An improved method, which made allowance for variability among variances of groups of students of different levels within the same secondary school grade, was used by

[3] Flanagan, John C. *Scaled Scores.* New York: The Cooperative Test Service of The American Council on Education, 1939. P. 33.

Flanagan in the development of his Scaled Scores for the Cooperative Test Service. Flanagan obtained his units by hypothesizing a number of overlapping normal distributions at different levels of ability in the range covered by the raw scores.

It is possible that normality of trait distribution is sometimes assumed in the definition of units when such an assumption is unwarranted. This possibility seems probable when one is considering the distribution of academic talent within a particular grade. Consequently, I developed a method for establishing units of measurement, called K-units, which produce interval scales in academic traits over a range of many grades when the trait under consideration is assumed to be distributed within each grade as a Pearson Type III Curve. Since the normal curve is a Type III Curve with zero skewness, such an assumption will lead to a normal curve if the data warrant it.

In a paper as brief as this no attempt will be made to describe the statistical processes by which K-units were obtained. In general the scaling procedure involves the determination of the appropriate skewness and standard deviation for the distribution of a particular trait in each grade. These parameters are found by fitting a series of overlapping frequency curves in such a way that the proportions of individuals in consecutive grades who manifest similar attainment in a particular trait correspond to the proportions given by the original data.

Neither a system of scores nor the establishment of units can be justifiably determined by fiat; they should be rooted in concrete objective data. Since K-units were determined from overlapping frequency distributions, it is desirable to define the reference population and to describe the sample of this population used. Unless the normative group is clearly defined and specified, neither units nor a system of scores derived from it is likely to be meaningful to the test user.

The population in terms of which K-units were defined is the national elementary school population (Grades 2 through 9). The sample used in determining K-units was the one used by the World Book Company in determining the norms of the 1941 revision of the Stanford Achievement Tests. It consists of approximately 50,000 cases and is itself a twenty per cent random sample selected from about 300,000 pupils to whom the Stanford Achievement Test Form D was administered at the end of the school year in 1940. The sample appeared representative of the national elementary-school population with respect to sex, I.Q., age, and geographical location.

Although scales in terms of K-units were obtained for only four traits, arithmetic comprehension,

arithmetic reasoning, paragraph meaning, and word meaning, the method could be used to obtain a scale in terms of K-units for numerous other traits.

A K-unit is defined as one-seventh[4] the standard deviation of the national grade 5 frequency distribution in any trait where Pearson Type III Curves have been fitted to it and the adjacent grades in such a way that the proportion of cases in each grade exceeding each raw score is that found in the original data.

Since K-units are units of an interval scale, any arbitrary reference point could be selected without affecting the validity of the scale. However, a point of reference was desired which would be meaningful to the users of K-scores. The mean performance of children in the United States after completing the ninth grade was selected as reference point and assigned a K-score of 100.

Since K-units provide an interval scale, let us now examine: (1) the average growth in arithmetic computation, arithmetic reasoning, paragraph meaning, and word meaning from year to year; (2) the extent to which the variability in each of the four traits changes as the children progress through the grades; (3) the effect of their progress on the skewness of the grade distributions; and (4) whether the fitted curves approximate normal curves.

Although it is commonly believed that growth in specific subjects in the elementary and junior high schools is not constant from grade to grade, the objective verification of this belief has been difficult due to the lack of an interval scale extending over the range of grades. The differences in mean achievement (in terms of K-scores) of the successive grades of the sample used in this study on the four traits, measured by the Stanford Achievement Tests, are given in the third column of Tables I to IV. Since all pupils were tested at the end of the school year in June, these differences are indicative of the amount of growth in the trait measured in the particular grade listed.[5]

The relative change in variability of the performance of children in successive grades has heretofore also been difficult to determine, due to

[4] Most tests, to serve their function adequately, should have a reliability in the neighborhood of .8 among the group to which they are administered. One-third the standard error of measurement for this reliability is approximately one-seventh sigma.

[5] However, since the people in each grade were different from those in other grades, these differences in grade means may be considered as growth only to the extent that we are willing to consider, for example, the present third graders as comparable to what the second graders will become a year hence. True growth could be determined by measuring the same people with comparable instruments in terms of K-units at different grade levels as they progress through school.

the lack of an interval scale extending over the range of grades. The standard deviation in terms of K-units of each grade in each trait is given by the fourth column of Tables I through IV. Table I shows a *consistent increase* in variability in arithmetic computation from the second grade to the ninth with the exception of one slight reversal occurring between grades 6 and 7. The standard deviation in grade 2 is 1.300 K-units while in grade 9 it has increased to 16.032 K-units.

Similarly in Table II it can be seen that there is a *consistent increase* in variability in arithmetic reasoning from the second grade to the ninth. The standard deviation in grade 2 is 3.165 K-units, while in grade 9 it has increased to 11.506 K-units. Thus, it can be seen that as children progress through the grades the problems of the arithmetic teacher increase in that the groups become more heterogeneous.

In contrast to the two arithmetic traits, Table III shows that the variabilities of the various grades in paragraph meaning are *nearly constant*. The standard deviation of each grade is of about the same magnitude (i.e., 7) as the standard deviation of the basic grade.

With the omission of grades 2 and 3, it can be said that the variabilities of the various grades in word meaning, as in paragraph meaning, are *nearly constant*. The standard deviation of each grade again is about the same as the standard deviation of the basic grade. The large standard deviations of 13.573 in grade 2 and 11.468 in grade 3 in word meaning may be due to inequality in home environment.

Undoubtedly families differ greatly in the degree to which they read stories to their children, engage them in family conversation, or otherwise stimulate their verbal development. A reasonable interpretation is that upon entrance to school children differ considerably in vocabulary. By the time the children have reached the end of grade 4, the school has tended to reduce the initial variability and make the group more homogeneous in its knowledge of vocabulary. In succeeding grades, the teaching is of such nature that no great change in variability occurs.

The skewness of each grade distribution for each trait is given by the last column in Tables I through IV. No consistent trends comparable to those observed for grade standard deviations are evidenced. The examination of these tables would lead one to believe that the assumption of normality for every grade distribution in each trait is not as tenable a hypothesis as the assumption that the grade distributions are skewed. There are, however, a large number of grade distributions indicated in Tables I through IV which are so slightly skewed that the assumption

of normality for these grades would not be unreasonable.

At the beginning of this paper it was emphasized that the laws of addition and subtraction were applicable only to scores constituting an interval scale. It was also pointed out that, rigorously, means, standard deviations, and correlation coefficients should not be computed from scores offering only relative order of data, because such statistics would be in error to the extent that the successive intervals on the scale are actually (not just numerically) unequal in size. Therefore, for group analyses involving the computation of means, standard deviations, and correlation coefficients, it is anticipated that educational workers may find K-scores more precise and useful than other types of scores. An additional study is now in progress to observe to what extent the trends observed in this paper are duplicated when a local sampling of grade performance is reported in terms of K-scores. Of course, it is in the actual application in various practical situations that the real worth of any new scale unit will be demonstrated.

TABLE I. K-Score Mean, Standard Deviation, and Skewness for Each Grade at End of School Year.

ARITHMETIC COMPUTATION

Grade	*Mean K-Score*	*Difference*	*Standard Deviation*	*Skewness*
9	100.00		16.052	.45
		3.44		
8	96.56		14.250	.40
		8.43		
7	88.13		11.001	.28
		4.94		
6	83.19		11.595	.82
		8.66		
5	74.53		7.000	.81
		6.36		
4	68.17		5.258	.58
		3.44		
3	64.73		5.142	.13
		4.83		
2	59.90		1.300	.41

TABLE II. K-Score Mean, Standard Deviation, and Skewness for Each Grade at End of School Year.

ARITHMETIC REASONING

Grade	*Mean K-Score*	*Difference*	*Standard Deviation*	*Skewness*
9	100.00		11.506	.73
		2.15		
8	97.85		11.192	.86
		6.55		
7	91.30		8.816	.85
		4.97		
6	86.33		8.798	.38
		5.85		
5	80.48		7.000	.53
		5.18		
4	75.30		4.937	.34
		4.87		
3	70.43		4.109	—.06
		4.45		
2	65.98		3.165	.16

TABLE III. K-Score Mean, Standard Deviation, and Skewness for Each Grade at End of School Year.

PARAGRAPH MEANING

Grade	*Mean K-Score*	*Difference*	*Standard Deviation*	*Skewness*
9	100.00		7.986	.29
		3.69		
8	96.31		7.457	.44
		3.37		
7	92.94		6.969	.50
		3.77		
6	89.17		7.152	—.39
		4.08		
5	85.09		7.000	—.10
		4.82		
4	80.27		6.665	.18
		6.45		
3	73.82		7.084	—.40
		6.65		
2	67.17		6.433	.31

TABLE IV. K-Score Mean, Standard Deviation, and Skewness for Each Grade at End of School Year.

WORD MEANING

Grade	*Mean K-Score*	*Difference*	*Standard Deviation*	*Skewness*
9	100.00		6.796	.46
		3.30		
8	96.70		6.191	.52
		2.84		
7	93.86		5.773	.66
		4.77		
6	89.09		7.103	—.27
		4.05		
5	85.04		7.000	—.04
		4.97		
4	80.07		6.824	.18
		9.05		
3	71.02		11.468	—1.04
		12.43		
2	58.59		13.573	—.50

Norms and the Individual Pupil

by LEE J. CRONBACH

Achievement tests are most often used to evaluate the educational development of a pupil. The teacher or counselor who refers to a norm table receives insufficient help, because the table does not give an adequate answer to the question, "Is this pupil doing as well as we should expect of him?" Often, the teacher must stop with the judgment that John is doing superior work, and that, since John is a superior student, everything must be in order. But John might be much less above average than his ability permits. And Sam, who is a below average student, may be dismissed without further thought, whereas a careful study of the evidence might show that he was really performing "over his head," or falling below reasonable expectation for such a sub-average student.

I would propose that we consider practical methods of providing achievement norms for different types of students. For some purposes, we need merely to compare the pupil with a large and representative group of the same age or grade who have had comparable experience. But usually we wish to decide if he is doing as well as others *like himself*. The teacher has been able, in the past, to look at the results for his class and, if they fell below the tabled norms, to stop with the easy evaluation that his group was weak to start with. If he had measures showing just how weak his pupils were initially, and we could supply evidence of how such weak pupils have performed in the standardizing groups, we would help him and impel him to think far more rigorously about the adequacy of the educational program he was operating.

This suggestion boils down to the following procedures. One or more reference tests would be given. As reference we might use a measure of mental age, of general educational development in the field, or a pretest of the sort we are norming. Then the evaluating test would be given. We would prepare a scatter diagram relating the evaluating test, on the Y-axis, to the reference test, on the X-axis. By plotting best-fit lines through the median of each vertical array, through the ninetieth percentile of each array, and so on, we would be able to provide a simple chart of norms. The teacher would interpret the performance of a pupil by

entering the chart with his X-score and reading his standing on the achievement test in comparison to others equal to him on the reference test.

This is not especially complicated for the teacher. Trimble and I[1] worked on methods of this sort with pretest-posttest data for certain P.E.A. tests and found that such graphic norms are readily interpreted. The proposal does involve additional work for the test standardizer, but in many cases the essential data are available in his files and the analysis can be done with little cost.

The first question we need to consider is: What shall we use as a reference test? The general answer would be: any variable likely to be correlated with performance. We already apply the principle of differentiated norms when we prepare separate norms for sex, age, or grade groups. Ordinarily, the most helpful reference test would be a measure of mental age or a pretest of the behavior being evaluated. When several tests are given simultaneously, as in a fall testing program, performance in each ability measured might well be referred against "general mental ability" or "verbal ability" as a reference. The producers of the Differential Abilities Battery, for example, have normed all their tests on the same sample and could easily produce charts of differentiated norms. I would suggest a broad search for helpful reference variables. Allison Davis' studies suggest strongly that separate norms for pupils of various social-class levels would stimulate teachers to do better teaching. If a pupil, although not outstanding according to general norms, is found to be quite superior considering his impoverished social background, a teacher is far more likely to put her heart into stimulating him. Other variables may prove equally significant for reference purposes.

This proposal does not conflict with the very desirable current trend toward providing separate norms for various types of schools or various curricula. But there are some interesting studies of how various types of pupils profit from different modes of instruction. They show that a particular change of method may lower the learning expectancy of one pupil while raising that of another. So there is a strong argument for norms by types of pupils within norms differentiated by curricula.

For what tests should we differentiate norms? In general, whenever there is a significant correlation between an achievement test and a sensible reference variable. In fact, I would go further and provide differentiated norms for all

[1] H. C. Trimble and L. J. Cronbach, "A Practical Procedure for the Rigorous Interpretation of Test-Retest Scores in Terms of Pupil Growth." Journal of Educational Research, 1943, 36:481-488.

guidance tests. One interprets mechanical aptitudes, for example, in relation to general academic ability, when making recommendations to a boy. Differentiated norms would reduce the subjectivity of such pattern interpretation.

Let us consider the relation of this proposal to the popular profile method. Profiles are becoming more useful as we get more sets of tests with norms based on comparable groups. But profiles are easily misinterpreted because of the regression fallacy. The differentiated norm chart automatically takes care of the factor of regression, and in that way is able to avoid the pitfalls of the achievement quotient. And by emphasizing differential interpretation, I hope it would make test-designers more conscious of the need for high reliability. The interpretation of profiles is severely complicated when the tests in the profile are unequally reliable, because the profile of true scores differs from that of the raw scores. This source of error, too, is eliminated when the profile is based on regressed scores. Finally, I may refer to our common experience that a test profile changes markedly when a different norm group is used. It appears sound to contend that the most meaningful profile is that derived by comparing the pupil with others like himself.

I recognize that I am passing over endless technical complications involved in this proposal. Some of these are: How can you standardize when every publisher will use a different test as a reference variable? How do you make a sensible reference variable out of such a complex mixture as general mental ability? How can you keep the system within bounds, when there are several possible reference variables for a given achievement test? How can you use growth norms, when the time interval between tests varies from school to school? And so on. But the main point is that we are giving test-users only a portion of the information that would be valuable to them, and that we could give them a lot more if we tried, with little extra cost. I wouldn't worry too much if our next type of norms is still only approximately utopian.

My concern about differential norms has been enhanced through recent experience with the Rorschach test. That, of course, is a test where norms are in an exceedingly primitive stage and where we have little cake of custom to hallow what norms exist. I have been using the test with superior adolescents and adults: people with I.Q.'s in the range of 120 to 160. When I try to make sense of a given score, all that is available to me is some investigator's finding that the normal person of a given age gives, say, 35 to 50 per cent F responses. Now I find my cases giving 60 to 80 per cent F responses.

What do I conclude? Only that they are different from the norm group. But I knew that to start with. The question I need to answer is whether my subjects differ, in the emotional characteristics the Rorschach reveals, from other people of the same high intellectual ability. There is no solution save to go out and prepare separate norms for a representative sample of highly superior individuals. While the Rorschach is certainly off our beaten track, I suspect that the same problem, in less striking form, is handicapping interpretation of much more prosaic tests of mathematics, interpretation of literature, and social attitudes.

Norms and the Individual Community

by ROBERT L. THORNDIKE

I am undertaking today, as my contribution to this symposium, to describe for you a project in which I am currently engaged and to tell you the values which I see in it.

In 1945, the World Book Company embarked upon a very ambitious program to establish norms for a new revised form of the Metropolitan Achievement Test. In the course of this enterprise, the tests were given to something approaching a half million children in some three hundred school systems spread over the whole United States. Each school system which participated undertook to administer the tests to all the children in at least three adjacent grades. A number tested all grades from the second through the ninth. A random sample of approximately 20 to 25 per cent of the pupils from each system were included in the final tabulations for establishing norms. In addition to the achievement battery, the appropriate level of the Pintner intelligence test series was given to each group.

The analyses of the above testing provided a mean score for each grade in each community for the intelligence test and for each of the subtests of the achievement battery. Translating each grade mean into a standard score for that grade, combining subtests of the achievement test and then combining data for the several available grades, a single standard score has been obtained to indicate the intelligence test level and achievement test level of each community. One then has an objective index of intelligence level and achievement level for each community in terms of which the communities may be compared.

I am currently gathering a number of economic, sociological, and educational facts about each of these communities. Most of the data have been taken from the reports of the 1940 census, but I hope also to get material from the published reports and unpublished statistics of the U. S. Office of Education. These facts I propose to relate to the criteria of intelligence and achievement for the communities. Working with the complete table of validities and intercorrelations, a study will be made to determine which combination of two, three, or more community variables gives the best prediction of the intelligence and achievement of the children of the community.

I see three types of values resulting from this study. The first is in connection with future enterprises of setting up test norms. If we can find a few community attributes which give a substantial prediction of community intelligence and achievement, it will be possible in the future, by selecting a sample of communities which has been stratified with respect to these characteristics, to achieve more representative norming populations and to achieve them a good deal more efficiently. Either the set of communities could initially be chosen in terms of significant variables or the set could be checked to see that they were representative and could be corrected for any bias that was found. It should be possible to use a smaller sample of communities and still obtain a more representative norming population and one for which comparability could be maintained from test to test.

A second value lies in the possibility of a more meaningful interpretation of national norms for the individual community. If two or three quite predictive community variables can be identified, then it may be possible to prepare a simple correction formula which can be applied to national norms to adjust them to the single community. This correction formula would permit the Superintendent or Director of Research to estimate what would normally be expected for a community having the characteristics of the one in which he works, and to appraise the performance of his school system in the light of this standard. He could even prepare an adjusted set of norms for his community.

A third value lies in the realm of pure knowledge. From the point of view of descriptive sociology it would seem well worth while to know what features of a community are most significantly related to the intellectual and academic level of its coming generation. It must be recognized, of course, that these relationships can not be assumed to indicate causation. But a first step toward understanding causal relations must be to determine the pattern of relationships which actually prevails.

This study is possible only through the generous cooperation of the World Book Company. Its initiation was stimulated by the sincere desire of Dr. Walter Durost, then Director of the Division of Research and Test Service for the World Book Company, that their rich store of data be used for research purposes.

Norms For Scholastic Aptitude and Achievement Tests of Independent Secondary School Pupils

by ARTHUR E. TRAXLER

It is well known that the interpretation of test results is greatly facilitated if norms are available not only for the general population but also for different groups within the total population. In the selection of clerical workers, for example, it is helpful for one to know not only how an individual's scores on a clerical test compare with those of persons in general, but also how they compare with those of persons employed in clerical positions. In the prediction of an individual's success in the study of accounting, it is of some help to know how he compares with high school graduates in general, but it is much more helpful to know how his scores on accounting aptitude tests compare with those of students entering their first course in accounting. In the consideration of an individual's fitness for admission to college in a certain geographical area, norms for that particular area are a useful supplement to national norms.

Similarly, in the guidance of elementary and secondary school pupils, counselors are greatly assisted if they have available norms for pupils whose backgrounds and whose educational and vocational aims are somewhat similar to those of the pupils they are advising. In particular, independent or private schools need norms for their own group of schools, since most of their pupils come from homes of an average, or better than average economic level and since all but a fraction of these pupils are preparing for college.

Extensive norms based on the test scores of independent-school pupils are available in the files of the Educational Records Bureau. For about twenty years, the independent-school members of the Bureau have been carrying on fall and spring testing programs with a wide variety of scholastic aptitude and achievement tests. Thus, dependable independent-school norms have been developed on many tests, and these are used as the basis of the Bureau's reports to schools and colleges. It is very important for those using these norms to be aware of the difference between independent-school norms and norms

for public-school pupils. Otherwise, erroneous conclusions may be drawn and marked injustice may be done to individual pupils.

The purpose of this paper is to present a small amount of evidence concerning the difference between independent-school norms and norms for less highly selected populations. It is difficult to choose from the mass of data in the Bureau files pertinent to this topic a small amount of illustrative material that can be included in a relatively brief presentation. But it is believed that what has been chosen is fairly representative of a much larger amount of material that could be drawn upon.

SCHOLASTIC APTITUDE

The distributions of total scores made by more than fifteen thousand independent-school pupils in Grades IX through XII on the American Council Psychological Examination, 1947 College Freshman Edition, are shown in Table I. The college freshman form rather than the high school form has been employed in the independent-school testing program because experimentation has indicated that the high school test is too easy for the more able pupils in the upper grades of the independent schools.

The preliminary college freshman median is shown by the broken line across the page. The independent-school medians for the different grades are indicated by the short horizontal lines adjacent to the distributions, and the range of the middle 50 per cent of the cases is shown by the vertical line parallel to each distribution.

It will be observed that the college freshman median falls between the independent-school medians for Grades X and XI. This kind of result has been obtained year after year in the independent-school testing programs. It will also be noted that the median for the entire group of independent secondary school pupils is slightly above the college freshman median. In other words, the independent-school population, including pupils in all four secondary-school grades, is equivalent in ability to the students entering the freshman year of college throughout the United States. The median for Grade XI is considerably above the college freshman median, and nearly three-fourths of the independent-school pupils in Grade XII reach or surpass the median for the freshmen entering college. These data do not mean, of course, that the average pupil in Grade XI or Grade XII of the independent schools is superior to the average student entering the colleges that select their students highly, but it does mean that eleventh- and twelfth-grade pupils in the independent schools, on the average, are more capable of undertaking college study than is the usual freshman in the group of two or three hundred colleges using the American Council

Psychological Examination, even though these pupils are still a considerable distance below the point of graduation from secondary school.

When one notes that the median total score for the independent-school twelfth-grade pupils is approximately fifteen points above the median for college freshmen, it is apparent that a pupil whose result is reported to a college on the basis of the norms for independent-school pupils at the twelfth-grade level will have a significantly lower percentile rating than he would have received if his result had been interpreted in terms of college freshman norms, even though the college freshman students are, on the average, a year older than the independent-school pupils in Grade XII. In fact, a pupil whose score is just at the median or fiftieth percentile for the independent-school twelfth-grade group would receive a percentile rank of 72, if his result were evaluated on the basis of the preliminary college freshman norms. One whose score exceeds those of only one-fourth of the twelfth-grade pupils in the independent schools is at the median or average for entering college freshmen. There are no norms on this edition of the psychological examination for public-school pupils, but it is logical to infer that the superiority of the independent-school pupils to the public-school pupils would be considerably greater than the margin of advantage which the independent-school pupils have over college freshmen.

The distributions of the median scores of the independent-school pupils in Grades IX to XII inclusive on the Iowa Silent Reading Test, Advanced Form Bm, are shown in Table II. These tests were administered in the fall. The end-of-year public-school medians for the four grades are indicated by the dotted lines across the distribution columns, and the college freshman median is shown by the heavier broken line on the distribution for Grade XII.

More than three-fourths of the independent-school pupils in Grades X, XI, and XII and nearly three-fourths of those in Grade IX have reading scores above the end-of-year public-school median for their grade level. The median for Grade X in the independent schools is a little above the public-school median for Grade XII. The eleventh-grade group in the independent schools has a median almost up to the college freshman norm, and the twelfth-grade median is considerably above this norm.

At the twelfth-grade level, a pupil whose reading score is at the median or fiftieth percentile for the independent-school group would have a percentile of approximately 83 on the public-school basis. One who reaches a percentile of 75 in the independent-school twelfth-grade group would be at a percen-

tile of 93 for public-school pupils in this grade. A pupil who is below four-fifths of the independent-school group in reading ability as measured by this test would be up to average for the public-school group.

A fair, general statement on the basis of the reading test data is that the independent-school norms for any grade at the high-school level are similar to those for public-school pupils two grades higher.

ACHIEVEMENT ON COOPERATIVE TESTS

It was necessary to choose from the large number of Cooperative tests used in the spring program of the independent schools certain ones which it was thought would represent best the independent-school achievement as compared with that in public schools. The independent-school results are based on testing in April, or, in the case of the science tests, March testing, while the public-school medians are end-of-year medians.

English.—The distributions of total Scaled Scores made on the Cooperative English Test by pupils in Grades VII through XII in the spring, 1947, testing program are shown in Table III. The general conclusions indicated by the English test results are not greatly different from those based on the Iowa Silent Reading Test.

More than three-fourths of the independent-school pupils have total English scores above the end-of-year public-school median for their grade. The independent-school medians in Grades VII through X are almost, but not quite, up to the public-school norms for pupils two grades higher. An independent-school twelfth-grade pupil who exceeds only 25 per cent of his group will be a little above the median for public-school pupils finishing Grade XII.

Mathematics.—The results of the Cooperative Tests in plane geometry, solid geometry, and trigonometry were chosen as representative of the comparative achievement in independent schools and in public schools in the field of mathematics. The distributions of independent-school scores on these tests are shown in Table IV. It will be observed that the median Scaled Score for tenth-grade pupils studying plane geometry is above the end-of-year public-school median but that the difference is not very large. In Grade XI, however, approximately three-fourths of the pupils have scores up to or above the median for the public-school group.

The median Scaled Scores made by independent-school twelfth-grade students of solid geometry and trigonometry are above the public-school norms for one-half year of study, but the differences are comparatively small.

The independent-school pupils surpass the public-school pupils in these subjects, but their superiority

is definitely less than in the preceding tests. There are two reasons for the fact that the independent-school pupils do not stand out as much as they did on the other tests. One is that, on the whole, independent-school pupils as groups regularly show greater superiority to public-school norms in the verbal field than in the numerical field. The other reason is that, so far as solid geometry and trigonometry are concerned, the pupils in public schools who take these subjects are rather highly selected. In other words, only those public-school pupils who do well in the lower courses in mathematics tend to attempt solid geometry or trigonometry. In the independent schools, on the other hand, it is quite common for pupils to take four years of mathematics. The point is that in these two subjects both the independent-school and public-school groups are selected. The independent-school group is, however, apparently still a little more highly selected than the public-school group studying these subjects. A somewhat similar situation with regard to selection will be noted in connection with the French test.

French.—Distributions of total Scaled Scores of independent-school pupils on the Cooperative French Test, Advanced Form, are shown in Table V. It is common for pupils in the independent schools to start the study of French and Latin in Grades VII or VIII or even lower. Thus, many of the pupils have had two or more years of study by the time they reach Grade IX. In order to keep the data comparable with those from public schools, however, the scores of independent-school pupils who began the study of French below Grade IX have been excluded from this report. Also, the scores of pupils who were in their first year of study are not reported here, for these pupils took the elementary form of the test.

The comparisons in Table V are between the distribution for the independent school tenth-grade pupils who were in their second year of the study of French and the end-of-year public-school median for two years of study, the distribution for the third year of study at the eleventh-grade level and the public-school median for three years of study, and the distribution for four years of study at the twelfth-grade level and the public-school median for four full years of French.

The independent-school medians are significantly above the public-school norms, but, just as was true of mathematics, the superiority of the independent-school pupils is not as great in this field as it is in reading and English. About two-thirds of the independent-school pupils have Scaled Scores approximately at or above the public-school norm.

One reason for the fact that there is somewhat closer correspondence between the independent- and pub-

lic-school results in French or any other foreign language than in English is that the study of foreign languages in itself is a selective agent in the public schools. On the whole, the more able public-school pupils tend to elect foreign languages and to continue with the study of them in the more advanced years. In the case of the independent-school population, however, nearly all the pupils take languages, for they are essentially a college preparatory group, and most colleges still require candidates for admission to submit a number of units in the language field.

Notwithstanding the selective nature of languages in the public schools, there is still a considerable amount of difference between the percentile norms based on independent-school and public-school results in French, A pupil at the median for the independent-school group equals or surpasses nearly two-thirds of the public-school group which has had a comparable amount of study.

Social Studies.—Distributions of Scaled Scores made by independent-school groups in April, 1947, on the Cooperative Tests in American history, ancient history, modern European history, world history, and economics are compared with end-of-year public-school medians in Table VI. It will be seen that the independent-school medians are definitely above the public-school medians. Except in world history, the independent school Q1, or twenty-fifth percentile, is a little below the public-school median. The difference between the independent-school and public-school norms on these tests is enough to affect significantly the interpretation of the percentiles. For instance, on the American history test, a pupil who is at the median for the independent-school group equals or exceeds 80 per cent or four-fifths of the public-school pupils in the East, Mid-West, and West whose scores contributed to the norms. One who is at only the twenty-fifth percentile for the independent-school group is not far below the median for public schools.

Natural Sciences.—It was necessary to go back to the 1940 Achievement Testing Program in order to get comparable data on the science tests, since in recent years special tests for which there are no public-school norms have been used in the independent schools to test achievement in biology, chemistry, and physics. The distributions of Scaled Scores of independent-school pupils on the Cooperative Tests in biology, chemistry, physics, and general science are shown in comparison with the public-school medians in Table VII.

On all four tests, the independent-school medians are much above those for the public schools. In biology and chemistry, well over three-fourths of the pupils surpass

the public-school norms, while in physics and general science a little less than three-fourths reach or exceed the medians based on public-school results. In biology, chemistry, and physics, the independent-school medians, or fiftieth percentiles, are equivalent to public-school percentiles of 85, 80, and 75 respectively.

SUMMARY

On the American Council Psychological Examination, College Freshman Edition, the independent-school medians for Grades XI and XII consistently surpass the median for entering freshmen. The twenty-fifth percentile for the twelfth-grade distribution is fairly close to the college freshman median. In reading ability and English achievement, the independent secondary-school pupils are about two grades advanced over the public-school pupils. In foreign languages and mathematics, there is a considerable amount of selection in the public schools themselves, and the differences between the independent-school and public-school norms, although still significant, are less than they are in the other fields. In natural sciences and social studies, about three-fourths of the independent-school pupils have achievement scores fairly near or above the public-school medians.

Although differences from field to field and from test to test make generalizations hazardous, the following general statements will usually apply fairly well to all areas except foreign languages and advanced mathematics. A pupil whose test score is equivalent to an independent-school percentile of 75 will usually obtain a percentile of 90 or better on the public-school basis. A pupil who is at the median or fiftieth percentile for the independent schools will usually be at the seventy-fifth percentile or higher, if his test result is compared with the public-school norms. A pupil who has a percentile rating of 25 on the independent-school basis is likely to be up to or above the public-school median. Pupils with independent-school percentiles as low as 10 or 15 are likely to fall within the middle 50 per cent of the public-school population although below the public-school median.

There is no indication in this paper concerning relative teaching efficiency in independent and public schools. The point of the discussion is that measurable differences do exist between the independent-school and public-school populations on nearly every test used by the Educational Records Bureau during a period of twenty years and that, where the test results are expressed in terms of independent-school norms, justice will be done to individuals by college admission officers, employers, and others only when the differences between the norms for these two groups are recognized and taken into consideration.

TABLE I

AMERICAN COUNCIL PSYCHOLOGICAL EXAMINATION, 1947 Edition

Distributions of Pupil Total Scores of Independent Schools in Fall Testing Program, September 15 to November 1, 1947

Score	*Grade 9*	*Grade 10*	*Grade 11*	*Grade 12*	*Total 9-12*
200					
195					
190					
185					
180				1	1
175				4	4
170			3	5	8
165		1	4	16	21
160		1	10	37	48
155		3	21	53	77
150	2	9	31	95	135
145	2	11	45	113	171
140	8	28	90	172	292
135	12	39	131	207	385
130	26	67	167	274	520
125	29	90	199	304	619
120	56	142	306	370	847
115	71	225	362	424	1067
110	114	220	408	432	1131
105	141	306	419	392	1231
100	179	335	425	371	1272
95	214	339	386	327	1231
90	251	340	334	267	1155
85	244	368	298	231	1143
80	261	305	239	146	934
75	210	252	187	126	826
70	200	183	152	81	626
65	158	160	84	55	499
60	128	107	61	45	371
55	94	65	45	24	262
50	68	51	23	9	177
45	42	39	18	11	136
40	23	25	15	4	86
35	12	9	8	2	42
30	4	7	3	2	24
25	1	4	4	2	14
20	3	2			5
15	2	1			4
10			1		5
5					
0-4					
Total	2555	3734	4474	4602	15365
Q3	100.0	113.4	123.5	132.1	121.7
Md	86.5	99.3	109.5	117.4	105.6
Q1	72.6	85.5	94.8	102.2	89.1
Range	16-151	23-171	16-177	32-189	16-189
Schools	106	124	126	126	132

Md = Median

-------- Preliminary college-freshman median

Maximum score is 200

TABLE II

IOWA SILENT READING TESTS, Revised New Edition, Advanced Form Bm

Distributions of Pupil Scores of Independent Schools in Fall Testing Program, September 15 to November 1, 1947

Score	*Grade 9*	*Grade 10*	*Grade 11*	*Grade 12*
216				1
214		1		5
212			7	24
210		4	20	66
208		14	32	44
206	2	11	18	53
204	3	8	10	29
202		23	63	117
200	4	28	67	99
198	3	26	39	61
196	15	63	125	163
194	16	33	62	89
192	20	75	130	194
190	31	90	151	153
188	37	69	147	139
186	49	78	146	186
184	40	82	142	110
182	44	68	87	64
180	134	227	274	269
178	71	92	104	74
176	76	121	143	105
174	144	201	175	150
172	126	126	136	89
170	115	90	81	89
168	249	257	213	57
166	44	50	39	148
164	154	153	113	23
162	138	91	61	83
160	68	47	30	25
158	111	77	47	22
156	49	23	99	23
154	94	45	30	8
152	43	34	27	20
150	51	39	13	5
148	48	21	9	12
146	32	16	5	3
144	20	6	5	5
142	22	9	5	1
140	19	9	2	2
138	6	5		
136	2			1
134	6	4		1
132			3	1
130	2	1	1	
128	1			
126	1	2		
124	2	2		
122			1	
120				
Total	2092	2421	2772	2724
Q3	176.6	184.0	190.4	195.6
Md	169.1	175.0	181.0	186.7
Q1	160.4	166.9	172.0	176.0
Range	124-207	125-215	122-213	135-127
Schools	86	79	79	74

Md = Median

..... End-of-year public-school median

---------- College-freshman median

TABLE III

TOTAL ENGLISH SCORE
COOPERATIVE ENGLISH TEST, FORM R

Distributions of Pupil Total Scaled Scores of Independent Schools Participating in Achievement Testing Program, April, 1947

Scaled Score	*Pupils Scores*					
	Grade 7	*Grade 8*	*Grade 9*	*Grade 10*	*Grade 11*	*Grade 12*
98						1
96					1	
94						1
92					1	6
90					2	4
88					2	10
86				2	2	9
84			1	2	5	18
82				3	10	17
80		1	1	5	9	20
78			1	6	8	24
76		1	2	6	29	42
74			8	11	25	41
72		1	12	21	35	72
70		4	16	25	50	110
68		4	31	31	65	67
66		2	27	46	96	99
64	2	10	52	62	89	130
62	2	18	44	70	109	101
60	2	16	74	91	148	147
58	1	23	92	122	163	172
56	3	29	86	137	170	173
54	15	49	108	116	182	127
52	8	49	97	157	146	126
50	12	70	122	152	138	87
48	19	44	112	104	89	69
46	20	57	108	121	110	60
44	21	76	108	79	53	34
42	37	87	105	64	43	20
40	32	50	77	54	35	13
38	39	59	88	29	14	7
36	30	38	45	24	9	7
34	15	27	42	19	9	3
32	26	18	24	10	4	1
30	16	9	17	8	3	1
28	5	7	11	2	1	
26	3	4	7	1	1	
24	1	1	1	1		
22		1	4			
20	1	2				
18-19		2				
Total	310	759	1523	1581	1856	1819
Q3	46.7	52.7	57.5	59.8	63.4	67.7
Md	41.2	46.0	50.2	53.6	57.1	60.1
Q1	36.7	40.9	43.2	47.7	51.3	54.4
Range	20-65	18-80	22-84	25-87	27-97	31-99
Schools	24	45	73	65	65	62

-------- End-of-year public-school median

TABLE IV

PLANE GEOMETRY, SOLID GEOMETRY, AND TRIGONOMETRY COOPERATIVE TESTS

Distributions of Scaled Scores of Independent-School Pupils Tested April, 1947

Scaled Score	*Plane Geometry, Form R* *Grade 10* *1 yr.*	*Grade 11* *1 yr.*	*Total* *1 yr.*	*Solid Geometry, Form P, Grade 12* *½ yr.*	*Trigonometry, Form U, Grade 12* *½ yr.*
90+					1
88		1	1		
86				1	2
84		1	1		
82	1	3	4	1	4
80	1	2	3	6	3
78	2	14	16	2	9
76	7	24	31	9	10
74	15	45	60	6	9
72	31	77	108	15	12
70	38	90	128	9	4
68	52	88	140	21	24
66	34	79	113	19	20
64	40	71	111	21	9
62	33	68	101	17	12
60	37	53	90	14	20
58	53	60	113	31	23
56	44	59	103	26	31
54	64	80	144	25	9
52	71	71	142	23	17
50	78	78	156	32	24
48	71	59	130	19	21
46	40	31	71	7	4
44	19	7	26	1	5
42	15	8	23		5
40	13	8	21		4
38	10	6	16		3
36	8	6	14		1
34	10	6	16		
32	8	8	16		1
30	8	5	13		
28	8	7	15		
26	5	3	8		
24					
22-23	6	3	9		
Total	822	1121	1943	305	287
Q3	64.8	69.5	68.1	67.3	68.5
Md	55.3	62.1	58.9	59.3	59.6
Q1	49.6	53.3	51.4	53.5	52.4
Range	23-83	23-88	23-88	45-86	32-92
Schools	43	39	79	23	24

Md = Plane Geometry, Form R, Grade 10, 1 yr. (between 56 and 54)
Md = Plane Geometry, Form R, Grade 11, 1 yr. (between 62 and 60)
Md = Plane Geometry, Form R, Total, 1 yr. (between 60 and 58)
Md = Solid Geometry, Form P, Grade 12, ½ yr. (between 62 and 60)
Md = Trigonometry, Form U, Grade 12, ½ yr. (between 60 and 58)

--------End-of-year public-school median

TABLE V

COOPERATIVE TEST, ADVANCED FORM Q
FRENCH

Distributions of Total Scaled Scores of Independent-School Pupils Tested April, 1947

Scaled Score	*Grade 10 2 years*	*Grade 11 3 years*	*Grade 12 4 years*
98			1
96		2	3
94		6	3
92	1	7	6
90	2	6	10
88	3	16	8
86	2	24	9
84	3	42	17
82	11	29	12 _Md
80	18	32	11
78	24	49	14
76	31	42	9
74	51	56 _Md	11
72	51	67	9
70	65	54	8
68	100	41	2
66	79	41	2
64	89 _Md	39	3
62	91	34	3
60	89	17	4
58	57	18	
56	66	10	
54	65	9	1
52	56	7	
50	31	6	
48	29	2	
46	23	3	
44	12	1	
42	16		
40	15	1	
38	10	1	
36	9		
34	1		
32	2		
30—	1		
Total	1103	662	146
Q3	69.7	79.9	986.8
Md	63.5	73.4	81.3
Q1	56.2	66.9	74.8
Range	29-93	38-96	54-98
Schools	70	65	29

-------- End-of-year public-school medians for first through fourth year of study

TABLE VI

COOPERATIVE SOCIAL STUDIES TESTS

Distributions of Pupil Scaled Scores of Independent Schools Participating in Achievement Testing Program, April, 1947

Scaled Score	*American History, Form T 10-12*	*Ancient History, Form P 9-12*	*Mod. Eur. History, Form Q 9-12*	*World History, Fm. 1937 9-12*	*Economics Form S*[1] *9-12*
98+	3				
96				1	
94	1			1	
92					
90	2			2	
88	2	1		2	
86			1	1	
84	3		7	7	
82	4		5		
80	4		7	3	
78	1	2	14	2	
76	11	8	19	5	
74	13	12	19	15	
72	36	16	32	13	
70	42	32	17	16	
68	85	30	24	23	
66	72	39	19	27	1
64	101	45	49	35	
62	71 _Md	49	78	33 _Md	
60	57	36	69	31	3
58	44	56 _Md	54 _Md	34	3
56	74	49	67	44	5
54	50	46	79	26	3
52	32	39	55	21	5
50	75	58	85	27	8
48	37	55	47	17	4
46	35	37	27	9	6
44	45	18	40	6	8
42	15	17	21	6	4
40	18	11	14	2	8
38	9	6	1	6	9
36	9	5			10 _Md
34	4	5	3	1	10
32				3	7
30	4				6
28		1			9
26					5
24					3
22					4
20					3
18					4
16					2
14					4
12					4
10					3
8					3
6					
4-5					5
Total	959	673	853	419	147
Q3	67.1	64.7	63.9	66.9	46.4
Md	61.0	57.6	57.6	60.5	36.7
Q1	51.7	50.5	51.4	54.5	27.5
Range	30-108	28-89	34-86	33-96	4-67
Schools	30	45	51	17	6

-------- End-of-year public-school median

[1] Raw score

TABLE VII

COOPERATIVE NATURAL SCIENCE TESTS, FORM Q

Distributions of Pupil Total Scaled Scores of Independent Schools Participating in Achievement Testing Program, March, 1940

Scaled Score	*Biology Gr. 9-12*	*Chemistry Gr. 10-12*	*Physics Gr. 10-12*	*General Science Gr. 8*	*General Science Gr. 9-12*
90			1		
88			1		
86			1		
84			4		
82	2	1	3		
80	1		6		1
78	3	7	4	1	
76	1	9	11	1	1
74	8	28	14		1
72	26	62	44	3	4
70	29	81	54	2	5
68	62	119	51	3	7
66	82	126 Md	96	4	14
64	112	157	65 Md	3	19
62	109	92	41	9	28
60	124 Md	104	41	6	36
58	101	63	43	7	39 Md
56	65	57	64	9 Md	29
54	50	29	82	8	25
52	66	24	41	1	16
50	52	15	40	7	17
48	18	8	20	6	31
46	27	11	12	4	10
44	31	6	6	2	13
42	14	1	4	4	2
40	13	2		4	14
38	7				11
36	5	1		1	9
34	3	2		1	3
32	1	1	1		
30	5			1	1
28	1		1	1	1
26				1	
24					2
22-23				1	
Total	1018	1006	751	90	339
Q3	65.2	68.9	68.3	62.4	61.7
Md	61.0	65.3	63.2	56.5	57.0
Q1	54.5	60.6	55.6	48.6	49.4
Range	28-82	33-82	29-90	22-79	24-80
Schools	51	49	44	8	22

--------End-of-year public-school median (one year of study)

Norms of Achievement By Schools

by E. F. LINDQUIST

What I have to say is not at all profound, and I should be able to say it in just two or three minutes. As elementary as is the observation I have to make, however, judging from current test standardization practices and from the literature on testing, it is one that is sorely in need of stressing even in as sophisticated a group as this one.

What I wish to do is to put in a plug for a special type of norm for educational achievement tests. It is a type of norm which, in consideration of the major uses generally made of achievement tests, should always be provided with tests of that character, but one which in actuality is very rarely provided.

Let me remind you first of two major uses which we generally claim for educational achievement tests. First, we claim that these tests will provide the teacher and counselor with a more adequate and dependable basis for the educational guidance of the individual pupil and for the individualization of instruction. Second, we claim that these tests will be useful to the school administrator as a partial basis for evaluating the content and methods of instruction, or in curriculum evaluation in general.

Now with reference to the first of these purposes, quite obviously we need norms that are descriptive of the distributions of measures of achievement of pupils. We interpret a pupil's score by comparing it with the scores of other pupils in the relatively homogeneous population to which he belongs. The second purpose, however, is the evaluation of *group*, not of individual, achievement. What the principal wants to know is, "How are we doing as a school?", or, "How does this grade group compare in *average* achievement in relation to other similar grade groups?" Quite obviously, what is needed here is a type of norm that is descriptive of distributions of measures of group achievement, i.e., of distributions of school averages, not of distributions of individual pupil scores.

Since such norms are rarely provided, what the school principal does is to attempt to interpret his grade averages or grade medians in terms of the norms that *are* provided—norms descriptive only of distributions of pupil scores. What he really does, without often knowing that he is doing so, is to assume that the school averages are distributed just as are individual pu-

pil scores. School averages, of course, are much less variable than pupil scores (although not as much less as many may assume) and he immediately makes wrong interpretations of his results on that account. A school average that coincides with the 75th percentile in the pupil score distribution, e.g., might be a remarkably high average, but, since it is "at the 75th percentile," it might not be regarded as high by the principal. What he does would not be so serious, however, if there were a constant relationship between the variability of school averages and the variability of pupil scores. In that case he could at least tell on what tests his school did well in relation to other schools, and could rank the tests in order of the relative performance of his school, even though he would be misled as to the true relative performance on any one test. Actually, however, the relationship is not constant, but varies widely from subject to subject and from one grade level to another.

I have with me here some data concerning the relative variability of individual pupil scores and of school means for 34 widely used achievement tests. I am not going to read all of these data to you but will just give you a sample of them. They are all based on large numbers of cases—in no instance upon less than 193 schools, in some cases as many as 400, and in no case less than 6000 pupils—so the results are quite stable so far as sampling errors are concerned.

The tests include tests of elementary school achievement, such as the Iowa Tests of Basic Skills, and tests of high school achievement, such as the Cooperative subject achievement tests.

In the first place, let me say that in every case the variability of school means is far larger than would be the variability of means of random samples of the same size drawn from the same total population. If, in any case, one were to make a simple analysis of the total variance into "between schools" and "within schools" components, one would find that the "between schools" variance is from two to ten times as large as that which could be accounted for by random sampling alone. The ratio would vary greatly, however, from test to test.

In third grade arithmetic (Iowa Tests of Basic Skills) the ratio of the 10th to the 90th percentile range for school averages and for pupil scores is exactly .5. For sixth grade arithmetic it jumps to .55, but for third grade language it is .36. For a high school physics test the ratio is .61, while for a U. S. history test at the same grade level it is .37. In general, the ratio seems to be largest where the opportunity to learn what is tested is relatively restricted to the classroom and to the formal school curriculum, and to be least where the achievement

measured might more often result from out-of-school experiences and informal learning.

Well, I think I have said enough to make my point. School averages are less variable than pupil scores, but much more variable than the means of random samples. More important, the relative variability of school averages differs markedly from test to test and grade to grade. If school administrators, then, are to interpret properly measures of *school* achievement for purposes of curriculum evaluation or for the evaluation of group achievement in general, they must be supplied with norms specifically adapted to that purpose—i.e., they must have norms descriptive of distributions of school *averages* for defined populations of schools, just as for the interpretation of individual achievement they now have norms descriptive of distributions of pupil scores for defined populations of pupils.

Some Basic Problems Concerning Achievement Test Norms at the Elementary and Intermediate Grades

by WALTER N. DUROST

I am very grateful to Dr. Thorndike for having explained to you about the standardization program for the Metropolitan Achievement Test, because that gives me an opportunity to omit that part of my paper and to devote just a half minute to saying that the original raw score data are still available at the World Book Company for the entire 500,000 cases tested, as well as IBM card samples for those cases used in the standardization program itself. Anyone else in the country qualified to use these data is certainly welcome to them, and I am sure that the World Book Company will cooperate to the fullest in making the data available in the form most useful for further analysis. For example, I think it would be excellent if the technique suggested by Dr. Gardner could be applied to the Metropolitan data. I am sure that the work that Dr. Cronbach has done would be further implemented if the Metropolitan data were used in connection with his analysis, and I am equally sure that the suggestions that Dr. Lindquist has made could be implemented by further analysis of the World Book Company data. And that brings me back to the basic problem of my own paper.

It is naively assumed by many users of achievement tests at the elementary and intermediate grades that all such tests distributed on a national basis are standardized on truly national populations. This is, of course, not true; no test, even at the present time, can claim that it is standardized on populations known to be *precisely* representative of the national scene. With the standardization of the Metropolitan Achievement Tests in 1946, the closest approach to date was made to this goal. In the Metropolitan standardization, a half-million children were tested, with every state being represented. All types of communities, varying in size from cities of 200,000 down to rural villages and county administrative units, were included in the program. Every community was tested in at least three consecutive grades and all children in those grades

were included in the testing program. Oddly enough, this program, far from simplifying the establishment of meaningful norms, has raised a number of basic problems which probably would never have come to the present stage of consideration if the program of standardization had not been carried out as indicated. Some of these problems I offer for your consideration in the few minutes at our disposal.

The sample on which the norms are based is large enough to permit fractionation into subgroups that will themselves still be large enough to be treated with some confidence. The question inevitably comes up as to how the sample might most meaningfully be broken down, and separate norms established for certain groups. As part of the standardization program, many types of related information about the communities involved were obtained. Intelligence test data, for example, are available for all pupils tested in all communities. Should norms be set up by IQ levels? Information is available as to the length of school term in each community. Should norms be set up based upon communities more or less homogeneous with respect to the number of days in attendance in a given school year? Some information is available concerning the curriculum in each of the cooperating communities. Should an attempt be made to group communities according to the progressiveness or conservatism of their courses of study? Also, there is ample evidence of tremendous differences with respect to retardation and acceleration policies, with respect to age of leaving school, with respect to amount of money spent per pupil for education. What about regional norms; separate norms for boys and girls; parochial-school norms; segregated negro-school norms—ad infinitum. There is, in a word, a large number of factors which are known or suspected to be related to achievement, any one or any combination of which may logically serve as a basis for differentiating the normative group into subgroups with the intent that normative groups so differentiated would constitute more homogeneous bases of comparison. An almost endless proliferation of norms is thus theoretically possible, and to some extent desirable—but one is very quickly confronted by the practical impossibility of providing such a multiplicity of norms in any economical, convenient way, and by the sad fact that the great majority of test-users would be confused rather than aided by this embarrassment of interpretative riches.

Here, then, is one question I pose for you. What are the criteria by which a testmaker may determine what normative groupings are necessary or desirable for a given test? It is platitudinous to say that the answer depends on the purposes for which the tests are to be

used; we need much more specific guides than this.

Apart from the problem of deciding for which groups norms ought to be developed, though not unrelated to it, is a second difficulty. This second problem stems from the fact that the so-called national norms for the Metropolitan tests, because they are based on a more nearly national sample, as well as because of real changes which have taken place in the schools in the past ten to twenty years, differ somewhat from previous norms for the old Metropolitan, the Stanford Achievement Test, or other achievement test batteries. In general, a lower average score is found to correspond to a given grade placement. It seems incontrovertible that this is a true reflection of the "national" picture, and that earlier "national" standardizations were based on samples of better-than-average achievement.

As the national norm is depressed, i.e., as a lower score is assigned to a given grade placement, there is an increase in the proportion of scores to which grade equivalents above 11.0+ (the highest grade equivalent given in the Metropolitan norms) are assigned. Thus, we have the difficulty that many communities in which testing is extensive, and in which the average IQ is substantially above the hundred mark, find not only that they are exceeding the national norm to a substantially greater extent than in previous testings with other batteries, or even with the earlier edition of the Metropolitan, but also that substantial numbers of their pupils at the 7th, 8th, and 9th grades obtain grade equivalents of 11.0+ and thus, for all practical purposes, seem to have undistributed scores. Since hardly any of the consumers of tests understand the procedure whereby grade equivalents are established and do not appreciate the problems involved in extrapolating norm lines, this is taken by them to be a defect of the test itself.

In a normal community, that is, one with an IQ of 100, one would expect that at the end of the 7th grade about 10 per cent of the children will get grade equivalents of 11.0+. When the average IQ goes up, or when the program of education in the community results in a higher level of achievement in any specific area, the proportion of children getting grade equivalents of 11.0+ increases substantially. It is not hard to see why a community using the Metropolitan and finding grade equivalents of 11.0+ for a large proportion of their children is inclined to complain.

One should hasten to make clear that the difficulty involved here has nothing to do with the amount of top in the test. If standard or even raw scores are used and distributions obtained for these children at any grade level, even for children substantially beyond the 100 IQ

mark, there is very little curtailment of scores at the upper end. The test itself is hard enough. The apparent curtailment is due to the use of an inappropriate method of interpreting the higher test scores.

What are reasonable solutions to problems of this nature? The following courses of action are suggested as appropriate:

1. If the populations involved are large enough, the establishment of local percentile norms is recommended. This permits the comparison of an individual pupil's performance with the larger population of which he is a part, which probably is a fairer basis for evaluating his performance than comparison with a national norm.

2. If comparisons are made with national groups, the users who are substantially above the norm must face the fact that their superior showing is a real fact and not due to inadequacies in the test norms, however disturbing they may find the truth in this particular case. Going back to old norms based upon different curricula in effect ten, fifteen, or twenty years ago, and based upon much less representative samples, is no solution.

3. If it is possible and within reasonable limits to do so, the publisher and authors may be willing to segregate sub-samples from the national population which comprise more reasonable groups for comparative purposes. Thus, it might be possible to segregate, for the sake of establishing more useful norms for independent schools, let us say, a group of communities well above 100 IQ and where the type of curriculum in effect, the length of school term, and other factors would be more nearly comparable. Percentile norms based on such samples would then be available for comparative purposes.

May I conclude by saying that it is indeed a strange dilemma when a publisher finds himself having problems with his test-using public because he has done an outstanding job from a technical point of view in establishing his norms on truly national samples. It points up the fact that one of the major issues facing the makers of tests, as well as the users, is a problem of education of the local test-user into a greater appreciation of the techniques involved in establishing norms and the limitations of these norms, particularly with regard to grade equivalents beyond the 8th or 9th grade level.

DISCUSSION

Participants:

JOHN C. FLANAGAN, E. F. LINDQUIST, ERIC F. GARDNER, WALTER N. DUROST, LEE J. CRONBACH, HUGH DAVISON, EDWARD E. CURETON, HAROLD GULLIKSEN, ROBERT L. EBEL, IRVING LORGE, DONALD A. PETERSON.

CHAIRMAN FLANAGAN: This concludes the original presentations of our six panel participants. Before proceeding to the discussion by the panel, I would like to raise one question which suggests a need for some cross-fertilization between the morning panels' ideas and some of the afternoon panel's ideas. It seems to me that a number of the people on the panel on validity this morning were talking about validity in terms of successful participation in a particular activity outside of the school situation. We had no suggestion, however, in our panel here this afternoon about something like successful participation norms based on groups who were successfully performing practical activities. This appears to be an area needing some further work. Maybe some of the panel members have something to say about this matter. Perhaps this is an indication of the discrepancy between the hypothetical ideas expressed this morning in the discussions on validity as contrasted with current practice as expressed this afternoon in the discussions of the norms that are available and are expected to be available in the near future.

I would like also to make another comment. Dr. Gardner's topic is one which is close to some of the work that I have done. Whereas many points are too technical for discussion here, the one general point which I think should be put into fairly sharp relief is that there are two different kinds of assumptions made in the procedures used for Scaled Scores and the procedures used for the K-units. In Scaled Scores we assumed normality within a single school system, in other words, within the group at Yonkers or within the group at Plainfield or Pittsburgh or some other city.

In Dr. Gardner's scaling work and that of many others, a form of distribution, such as a normal one or a skewed distribution of Pearson's Type III, has been assumed within some mixed sample of school systems representing a na-

tional, regional, or state group. It seems to me that it is much more reasonable to assume a stable distribution within one school system, where the length of school year, the textbook, the course of study and many other things are similar, than to assume stability for a mixed group containing 40 per cent New Englanders, 20 per cent from the South, and some other per cents from other regions. Similarly, a mixture of big cities and little cities does not appear likely to have a stable distribution form. If you take out one large city and put in another one you would usually change the general shape and form of the distribution. It appears likely that single school systems would be more stable in their distribution form. This doesn't necessarily mean that one system is not useful and the other one is very useful. The test of that is in the practical results that are obtained. I also think that Dr. Gardner has done some excellent work in his efforts to develop units for the particular problem that he had.

There are many, many problems of a technical nature connected with this type of scaling. The presentations are now open for discussion by the panel.

E. F. LINDQUIST: I would like to comment on Dr. Gardner's paper. It seems to me the principal trouble with his system of scaling is that it will lead to interpretations which are totally unacceptable to the school principal, or that are contrary to common sense. That probably wouldn't bother very many test technicians, but in this case I think it would be contrary even to the common sense of the test technician himself. I am speaking now about the application of this scale at the elementary school level with the type of elementary school achievement tests that Dr. Gardner used in establishing the scale.

If we scale a battery of educational tests in this fashion, we are in a sense assuming that the variability within the fifth grade—which happens to be the reference population here—is the same in all of the fields tested. That is not a tenable assumption, and I think I have some evidence here that will make that quite clear.

I suspected that this point would be raised, so I brought along with me some charts that I have had prepared for a course in measurement theory. These charts will indicate that the variability within grades in relation to grade differences differs so markedly from one grade level to another, or from one subject to another, that this system of scaling will lead to untenable interpretations.

This first chart presents frequency polygons representing distributions of scores of third, fourth, and fifth grade pupils on the read-

ing comprehension test of the Iowa Tests of Basic Skills. This second chart similarly shows the overlap in third, fourth, and fifth grade distributions on the arithmetic test of the same battery. The distributions are based on the scores of about 9000 pupils per grade. You will note a very marked difference in the degree of overlap from grade to grade for these two tests. The implications of this are clear. If one attempts to describe grade differences in terms of within-grade variability—on the assumption of constant within-grade variability from test to test—one would have to conclude that average growth in achievement from the third grade to the fifth is more than twice as great in arithmetic as it is in reading comprehension. Again, one would have to conclude that two years' growth is over four times as great in arithmetic at the third to fifth grade level as it is in reading comprehension at the sixth to eighth grade level.

These conclusions would be utterly unacceptable to the school principal. If anything, it is more plausible that more rapid growth takes place in reading than in arithmetic. The pupil's opportunities to achieve or to grow in arithmetic are far more severely restricted by the content and organization of, or the grade-placement in, the curriculum than they are in reading. The pupil's opportunities to learn arithmetic are much more confined to the classroom than is the case in reading, and the learning process is much more "lock step" in character.

DR. GARDNER: I would like to point out one thing, in reply to Dr. Lindquist's criticism, that these scales were constructed independently of each other, that is, in each case the technique was applied to one particular trait and it was assumed that the standard deviation of the basic grade, grade V, for that trait was 7. That does not imply necessarily that I believe that the relationship Dr. Lindquist mentioned holds. It seems to me that it might be possible to use some of the data that Dr. Lindquist has and see the relationship between the variability in my basic grade in one trait, the variability of the basic grade in another trait, and then there might be some possibility of comparing these; but there was no intention on my part to offer scales where there was comparability from one trait to another. These are independent scales.

CHAIRMAN FLANAGAN: I wonder if Dr. Lindquist has an answer to what the real relation is. How do you tell whether the variability is twice as great in one grade as another?

DR. LINDQUIST: I don't think we can tell. I think the relationship of within-grade variability to grade

differences is so largely determined by arbitrary factors in the organization of the school curriculum that it is just futile to attempt to establish a "true" scale on the basis either of the assumption of constant grade differences or of constant within-grade variability. As a practical solution to the problem, I think the assumption of constant grade differences is the more tenable, or more practicable, but only at the elementary school level. I don't think that is true at the high school level. There, I would prefer to use the type of scale that you have suggested.

CHAIRMAN FLANAGAN: There are two problems involved there. I think Dr. Gardner feels that his main contribution is a uniform type of unit, not the particular selection of unit size. What you are quarreling with is unit size and not his particular type of equality unit.

DR. LINDQUIST: Yes, but I consider the fundamental problem to be to secure comparability, and I suggest that we work on the fundamental problem first.

CHAIRMAN FLANAGAN: I can see what your point is, although one could still do something about equal units. I certainly agree with Dr. Lindquist that you are going to have units of particular size by definition and no other way. The 1/7 of the standard deviation of the fifth grade group as developed by Dr. Gardner, is a very clear and definite definition, and if some people find these units useful, that is fine; if other people do not find them useful for their purposes, they should use other units that are better suited to their problems.

WALTER N. DUROST: I would like to say that in the standardization of the revised Metropolitan test a scaling technique was used similar to that used on Stanford, except in this case we did set up scales which were independent from one subject to another, and we did make the assumption that, at the scaling grade which was the model age group at the sixth grade, the variability was the same. We established the mean at 200 and standard deviation of our scaled scores at 20 points, and then allowed the variability to change up and down the grade scale as it would in terms of our data.

Now, while we may be lifting ourselves by our bootstraps by assuming that the variability of scores is the same at one given grade point, it seems to me that even so that does give us some basis for finding out to what extent the variability does change from one grade to another, especially if we consider that the scales are independent of each other from subject to subject.

LEE J. CRONBACH: I don't think Dr. Durost and I are going to agree; maybe I will have Dr. Lind-

quist against me also. I don't think the grade equivalents are worth producing. I think they have, over the last thirty years, done more to confuse educators than any possible help. It seems to me if we threw them out the window and stuck to percentiles within grade groups or standard scores within grade groups (with some thinking about the retardation problem and the other things that do complicate all norms), we would have a much more intelligent system.

Now, I don't know what the arguments are on both sides here; it is a pretty long problem, but certainly you must realize that it is very foolish to say, for example, that some girl, who is in the sixth grade, is at ninth grade level in history—she hasn't had any ninth grade history—and seventh grade level in arithmetic. Maybe she is more outstanding in arithmetic in her group than she is in history.

DR. DUROST: I would like to say, in response, that I am entirely in agreement with the point of view presented, except I would like to say that the publishers do have to consider the test-using public. At that the responsibility may come eventually right back to this group, because the education of our teachers in the field of tests and measurements will not permit us to go over to percentile norms right now (for the elementary grades), but if we get busy and do something about teaching our teachers to use percentile norms, then we can make the shift.

The Metropolitan tests now provide percentile norms for all grade levels for three times in the school year and I would be very happy indeed if schools would accept percentile norms in place of grade equivalents, because I agree especially at the junior high school level and up, because, above those grades, grade equivalents are meaningless, if not positively vicious.

CHAIRMAN FLANAGAN: We will now open the discussion to participants from the floor.

HUGH DAVISON (Pennsylvania State College): One of the hottest issues in education today isn't quite so much how you make the examinations or even how valid they really are, but how well they please the parents. The parents are demanding, in many local situations, that the reading in the school be well-taught, and I single out reading as the center of all the arguments on educational method. What the parents want to do is to say that the reading in their school is at a level equivalent to the reading in some other school, and it practically forces us to use nationwide norms of some sort. The grade level seems to be about the best one we have. I think if we threw it away we might as well disregard at this point the idea that we are serving the public.

EDWARD E. CURETON: I would like to make some remarks on two or three of these points, I hope rather briefly. It is my impression, for instance, that the standard deviation of heights in inches is about four inches, and that the mean is about sixty-eight inches. Now, this mean is actually measured from the floor, and so we can point out that the mean is about seventeen times the standard deviation. In the case of weight, on the other hand, the standard deviation is on the order of twenty pounds, I imagine, and the mean is on the order of 150 pounds; the mean is about 7½ times the standard deviation instead of seventeen. I see no reason for supposing that a different situation exists in the case of mental traits and traits of achievement. In other words, it would not surprise me at all if reading is intrinsically twice as variable as arithmetic in a human population if we could measure it from a true zero point, and data such as Dr. Lindquist has presented could be interpreted to suggest that that might well be the case. That is, if the grade differences are several times as large in terms of the standard score in arithmetic, we might reverse the conclusion and say, not that people grow twice as fast in arithmetic, but, on the contrary, that variability in reading, which is subject to a much wider variety of influences than arithmetic, is twice as great.

DR. LINDQUIST: Evidently, then, you don't want to assume equal variability in arithmetic.

DR. CURETON: No. I am strongly against the assumption of equal variability without its actual demonstration, and I would feel happier with even grade units, though I think age units would probably be closer to the truth.

DR. LINDQUIST: I wish you would suggest how you would demonstrate it.

DR. CURETON: I leave that to the time at which we can all agree upon a technique like Dr. Gardner's or some other for the absolute scaling of tests and for the location of the true zero point of the measurements.

I should like to make another remark concerning Dr. Cronbach's presentation. If you have a posttest and a reference test which have a low correlation with each other, you can well question why you should interpret the posttest in terms of that reference test. If you have a posttest and a reference test which are highly correlated with each other, the difference between the two standard scores of one pupil will have a fantastically high standard error. I tested this some years ago in the case of the Stanford Achievement Test and the Stanford-Binet; if I remember the results correctly, the standard response error of the accomplishment quotient is about half of the stand-

ard deviation of the A. Q.'s themselves for these two fairly reliable tests. In other words, before you can be reasonably sure that an A. Q. is significantly different from zero, its numerical value has to be 15 or 16 points away from 100.

A final point regarding the matter of E. Q.'s and I. Q.'s: When you measure intelligence, whether by I. Q., mental age, or any other interpretive unit, you measure it by asking the examinee a number of questions and counting the number of right answers. When you attempt to measure his achievement, you do exactly the same thing. Therefore, I see no reason for assuming that the I. Q. is necessarily any nearer his heredity than is the E. Q.

We have commonly supposed that that is the case. Actually, the only way we can infer hereditary differences from test scores is by making the assumption that opportunities for learning and pressures leading to learning are homogeneous for the group measured. Now in the case of achievement tests, say the Stanford Achievement Test given at the end of the eighth grade in a particular school system, we have some reason for supposing that that is the case. This test contains items all of which measure, to a reasonable degree at least, something that has been taught. The school for eight years has exerted all sorts of pressures upon the pupil, and given him more or less unlimited opportunities to learn the things which the test then proceeds to measure. Therefore, if there is anything which satisfies the criterion for a test of heredity, it is a valid school achievement test given to a group of pupils who have all just finished the same grade.

I know of no comparable reason for supposing that a set of items measuring as wide a range of abilities as those measured, say, by the Binet, measure traits for which the individuals taking the test have had equal learning opportunities and equal learning pressures. So it seems to me that if you want to make any generalizations about heredity, you could say that the hereditary element in the achievement test is rather larger than the hereditary element in the intelligence test, and that what you are measuring with anything comparable to the A. Q. is the comparative uniformity of the social pressures in the community leading to mental development in areas other than those of the school subjects.

HAROLD GULLIKSEN: I would like to comment that it seems to me the problem posed by Dr. Lindquist's data has so far been discussed too much as if it were a technical test theory scaling problem. Discussants have asked, "Shall we assume equal variability in a given grade for various subjects, or shall we assume equal intergrade differences for various subjects?" I don't quite see how one assumption leads to any more acceptable inter-

pretations than the other. On one assumption you would say that there is very little growth from this grade to that, and on the other assumption you would say that the growth is all right, only many children don't make it. I would think that, if one interpretation is inacceptable to the school principal, the other is, also.

It seems to me the way to approach the problem is indicated by the remarks that Dr. Buros made this morning. I can't see that it is possible to speak in any meaningful sense about *the* grade variability or *the* grade means or *the* difference, and so on. Such statistics depend entirely on what items you put in your test. I suppose the children do learn something from the third to the fifth grade, and that they forget other things. The problem is, which set of items do you select to put in your test? Apparently in arithmetic we have selected items for the fifth grade that they didn't know in the third, and that they learn in the meantime. In reading, we have selected items for the fifth grade that they already knew in the third. It seems to me the problem, then, is in the closer liaison between the subject-matter expert and the test technician. If the conclusion that the children do not learn anything is inacceptable to the school principal, he ought to be able to present you with some items which do show learning, and then you can decide whether or not (or he can decide whether or not) you should use those items. If your tests seem to show too much learning, so to speak, then doubtless with due consideration you get other items that the children didn't learn too well in the meantime and put those in the test. I can't see this as a scaling problem at all; it is a problem of what items you initially included in your test, the relationship of those items to what was taught and what was supposed to be taught in the curriculum.

I come back to the topic Dr. Buros emphasized this morning, that is, you must—at least in part—give some recognition to what the teachers think they are teaching, and then if you present evidence that they haven't taught it, they shouldn't just throw that out of court and say, "This is inacceptable"; they should say, "We will have to improve our teaching methods." If you present evidence that the children haven't learned something that the teachers didn't think they were teaching anyhow, then the test should be changed. However, in either case the problem is one of test content in relation to curriculum and not a scaling problem.

ROBERT L. EBEL (State University of Iowa): I would like to comment on Dr. Gardner's paper with reference to the search for equal units. The possibility of equal units depends primarily

upon the thing being measured, and only secondarily upon the scale that is used. For measures of height, weight, and certain other things, it is possible to derive a scale on which the units can be shown to be equal by direct physical comparison. Further, the process of addition has operational meaning with such units. Neither of these statements can be made correctly, with reference to the units on the scale, of educational test scores.

It is of course possible to obtain a scale for educational test scores on which the units are *called* equal if they meet certain prescribed requirements. But there is a multiplicity of such possible requirements, and, since none of them can avoid the fact that fundamental demonstration of equality is impossible, there is no *intrinsic* reason why one set is better than another. The criterion of which should be used in any given instance is solely that of utility. In some cases it may be most useful to consider that the raw score units are equal. In others, the tester's purpose may be best served by devising some new type of unit such as the K-score.

Further, it would be wrong to assume that such arbitrarily defined equal units can be added in the same meaningful sense as inches or pounds can be added. It would be equally wrong to assume that they must never be added under any circumstances. Raw scores, z-scores, percentile ranks, and other types of scores may sometimes be added to produce useful new measures. But it should never be assumed that such addition has the same meaning as it has in physical measurement.

IRVING LORGE: I just want to say I think Dr. Gulliksen has made the most important suggestion of this afternoon's session, that is, the variability you get on the test is a function of the items you put into it and you can change the variability at will by choosing your items appropriately either on the basis of try-out data, or by judgment. The variability, therefore, is an inextricably confused variability, since different behavior is being evaluated within the individual, and, of course, among the individuals in the group.

Dr. Gulliksen made the point that he may have forgotten something he learned in the third grade. I like to ask my classes in Teachers College whether the Tropic of Capricorn is north or south of the Equator. I honestly do not know. I have been very careful not to reacquire that piece of information; otherwise, it would spoil the story. I did know that particular fact when I was in the fifth or sixth grade. I have lost many such items of information in the process of growing up and deteriorating, and, not only that, but the way I arrive at the answers to some questions on

achievement tests, or on intelligence tests, depends upon the process that I use. Some of the answers I know by memory; some I infer from relationships, and many I answer by eliminating all the wrong answers. The process must be considered in interpreting test scores.

Since a score of 34 right merely means 34 points regardless of which 34, or a particular 34 right rather than any other 34 right, the scores are equivalent merely in a counting sense. Manipulation of items, therefore, leads to different scores, and hence different notions about whether the variability of arithmetic in the third grade is greater or less than it is in the sixth grade. You are dealing with variabilities which are artifacts of the items and of the behaviors of the children that are being appraised.

I think we recognize that what we are dealing with is an indeterminate number, the meaning of which is rarely very clear except when referred to some group. At the best the numbers are averages in the same sense that the numbers in the study of gases are averages. The numbers do not deal with specific electrons but rather refer to a notion like average velocities. If we realize that our numbers have reference to a group and not to an individual we would be more careful in making some of the statements about relative variability.

DR. LINDQUIST: May I answer that? I am afraid I disagree with Dr. Lorge. I don't think Dr. Gulliksen's remarks or Dr. Lorge's remarks are very relevant to the issue.

Let me refer again to the test distributions that I showed you. The items in those tests were not selected because they showed grade growth or otherwise; they were not selected because they showed a high or low correlation with a test criterion. They could have been selected in that fashion, that is true. It would be very easy to build an arithmetic test that would show greater within-grade variability in relation to grade differences, or vice versa; but what we did in this case was to build an arithmetic test that defined, as well as we were able to define them, the things that teachers are trying to teach in arithmetic. After having built that kind of test, we then found this kind of relationship between within-grade variability and grade differences. It is beside the point to say that we could produce a different picture by changing the test content. The relationship is as it is, and it does make it very difficult, as a practical problem, to provide the schools with a system of scaling that will make results comparable from test to test.

DR. GARDNER: I also would like to make one comment in reply to Dr. Lorge. The technique that was

used in the determination of K-units did not depend upon shape of distribution of the original test items. It is true that naturally any result that we have is a function of the items themselves; but, in the technique which was used to develop K-scores, I accepted only rank order, and any test which would rank individuals in appropriate rank order would give me the same K-score units. So there is a little more than just the distribution of the raw scores involved.

DONALD A. PETERSON (Life Insurance Agency Management Association): I think the best way to settle the present discussion is to go on to another topic, so I would like to go back to Dr. Cronbach's discussion. In addition to the assumptions that have already been indicated here anent the remarks about similar ability curve or similar learning curve for people who have the same score on the reference test, it seems to me that there are one or two assumptions made, at least that is my impression.

These assumptions are either, that all those people who get the same reference test score are at the same point, equivalent points, on their individual learning curves, or, that their learning curve, or development curve, functions—whichever you want to call them—are linear. Now, to the extent that either of those assumptions is incorrect or unjustified—and I suspect that they are incorrect or unjustified—you are then going to get systematic error in the interpretation of individual cases. It won't affect group interpretations, such as how good a job the teacher is doing over a group, but it will introduce systematic error in the interpretation of individual cases. Would you comment on the correctness or incorrectness of those impressions of mine?

DR. CRONBACH: While responding to that, I think I should also refer to Dr. Cureton's remarks. I don't think we have any fundamental disagreement. I think the points made about reliability and definition of variables are quite relevant, but they are also found commonly in the sorts of interpretations teachers are making of tests, and I think this calls attention to them.

I am not completely clear on your point, Don. The assumption, I think, that we do conventionally make is that people with equal scores are at equal points on the learning curve. Then if we find during the year that this particular person has made greater growth than most other people, that certainly shows flatly that he is at a different point, that his learning curve is steeper. That is essentially the interpretation we are making.

Have I missed your point?

MR. PETERSON: I think you are all wrong. We have only a slight disagreement.

Let's take two persons, and here are two curves like that, and like this. If you take this as the developmental period along this line here, these two persons, even though they are at the same point with regard to your reference test which is scaled on the Y-axis, are not at equivalent points in their learning curves, and the normal development for the one as against the other over a certain period is simply not going to be the same. So that when you say, given this, the person should then be here, that just isn't true unless you can make the assumption that both individuals who get the same reference test score are at the same point on their learning or development curve.

DR. CRONBACH: I am not sure I would follow your use of the words "should be" at this point.

MR. PETERSON: I shouldn't have used those words; I'm sorry.

DR. CRONBACH: We could say literally this person has gained more than most similar people do. Does your argument hold if you use a pretest for your reference test?

MR. PETERSON: Oh, yes. I should have used two similar curves, one person starting here, one person starting here.

DR. CRONBACH: You put him on the program next year and let him develop what we ought to do.

CHAIRMAN FLANAGAN: I think we had better conclude the program.

Appendix

INVITATIONAL CONFERENCE ON TESTING PROBLEMS

Participants

ABEL, Walter K., Stevens Hoboken Academy
ACKERMANN, Jean M., Educational Testing Service
ADAMS, Elizabeth C., Educational Testing Service
AFFLERBACH, Janet G., Educational Testing Service
ANASTASI, Anne, Fordham University
ANGELL, George W., Michigan State College
ARNOLD, Samuel T., Brown University
ARSENIAN, Seth, Springfield College

BAGGALEY, Andrew R., Harvard University
BARKAN, Arlene E., Educational Testing Service
BARNES, Harry G., Educational Testing Service
BARRETT, Dorothy M., Hunter College
BEAR, Robert M., Dartmouth College
BECK, Hubert P., The City College of New York
BERG, Harry D., Michigan State College
BERGESEN, B. E., Jr., Educational Testing Service
BESER, Naomi H., Educational Testing Service
BOROW, Henry, University of Minnesota
BOYLE, Maribelle, Wayne University
BRANDT, Hyman, Personnel Research Section (AGO)
BRANSFORD, Thomas L., Albany, New York
BRIDGES, Claude F., World Book Company
BROOKS, Richard B., College of William and Mary
BRUMMETT, Thelma L., Metropolitan Life Insurance Company
BRYAN, Miriam M., Educational Testing Service
BUCKINGHAM, Guy E., Allegheny College
BURKE, Paul, Educational Testing Service
BURNHAM, Paul S., Yale University
BUROS, Oscar K., Rutgers University
BURR, Emily T., Stevens Hoboken Academy
CARROL, John B., Department of the Army (AGO)
CARSTATER, Eugene D., Navy Department
CHAUNCEY, Henry, Educational Testing Service
CHURCHILL, Ruth D., Antioch College
CONRAD, Herbert S., U. S. Office of Education, Washington
COOK, Elaine Forsyth, Detroit, Michigan
COOKE, F. Morse, Wayne University
COPELAND, Herman A., Atlantic Refining Company
CORNELL, Ethel L., New York State Education Department
COUEY, Fred, Air University
COWLES, John T., Educational Testing Service
CRANE, Percy F., University of Maine
CRAWFORD, John E., Wagner College
CRAWLEY, S. L., The City College of New York
CRISSY, William J. E., Queens College
CRONBACH, Lee J., University of Illinois
CUMMINS, Edith, Jewish Employment and Counselling Service
CURETON, Edward E., University of Tennessee
CYNAMON, Manuel, Educational Testing Service

DAVIS, Frederick B., George Peabody College for Teachers
DAVISON, Hugh M., Pennsylvania State College
DETCHEN, Lily, Pennsylvania College for Women
DRAGOSITZ, Anna, Educational Testing Service
DRESSEL, Paul L., Michigan State College
DUROST, Walter N., Boston University
DUVALL, Ellen N., Medical College of Virginia
DYER, Henry S., Harvard University

EBEL, Robert L., State University of Iowa
ENGEL, John B., Educational Testing Service

ENGELHART, Max D., Chicago City Junior College

FEIFEL, Herman, Personnel Research Section (AGO)

FEINBERG, Robert M., Educational Testing Service

FERGUSON, Leonard W., Metropolitan Life Insurance Company

FIKE, Irene A., Scarsdale Public Schools

FINDLEY, Warren G., Educational Testing Service

FISHER, Kathryn B., Educational Testing Service

FLANAGAN, John C., University of Pittsburgh

FORLANO, George, South Ozone Park, New York

FREDERIKSEN, Norman, Educational Testing Service

FREEMAN, Paul M., Air University

FRENCH, Benjamin J., New York State Civil Service

FRENCH, John W., Educational Testing Service

GARDNER, Eric F., Syracuse University

GERBERICH, J. Raymond, University of Connecticut

GORDON, Hans C., Board of Education, Pennsylvania

GROSSNICKLE, Foster E., State Teachers College

GULLIKSEN, Harold, Educational Testing Service

HAGEN, Elizabeth P., Teachers College, Columbia University

HAGMAN, Elmer R., Board of Education, Connecticut

HAUSRATH, Alfred H., Navy Department

HESTON, Joseph C., DePauw University

HILDRETH, Gertrude, New York City

HOGAN, Ralph M., Navy Department

HORTON, Clark Willis, Dartmouth College

HOSMAN, Captain Richard S., The Air University

HOTELLING, Harold, University of North Carolina

HUGHES, John H., California Test Bureau

HURD, Archer W., Medical College of Virginia

HURD, Charles P., Columbia University

JACOBS, Robert, Educational Records Bureau

JANOVITZ, Olga M., Hunter College

JASPEN, Nathan, Pennsylvania State College

JONES, Edward S., University of Buffalo

JONES, Galen, U. S. Office of Education, Washington

KABACK, Goldie R., The City College of New York

KATZELL, Raymond A., Syracuse University

KEMBLE, Elizabeth L., National League of Nursing Education

KIRKPATRICK, Forrest H., Bethany College

KOGAN, Leonard S., Community Service Society

KURTZ, Albert K., Pennsylvania State College

KUSHNER, Rose E., The City College of New York

LANGMUIR, Charles R., Educational Testing Service and Syracuse University

LANNHOLM, Gerald V., Educational Testing Service

LAZO, Elizabeth K., American Public Health Association

LENNON, Roger T., World Book Company

LEVERETT, Hollis M., Northeastern University

LINCOLN, Abraham L., The Lawrenceville School

LINDQUIST, E. F., State University of Iowa

LONG, Lillian D., Merit System Unit

LONG, Louis, The City College of New York

LORD, Frederic M., Educational Testing Service

LORGE, Irving, Teachers College, Columbia University

LURIE, Walter A., National Community Relations Advisory Council

LUSK, Louis T., Norwalk, Connecticut

LUTES, Olin S., Brooklyn College

MACPHAIL, Andrew H., Brown University

MCCALL, William Charles, University of South Carolina

MCNAMARA, Walter J., International Business Machines Corporation

MCQUITTY, John V., University of Florida

MANUEL, Herschel T., University of Texas

MATHEWS, Chester Ora, Ohio Wesleyan University

MEYER, Lorenz A., Veterans Administration

MICHAEL, William Burton, Princeton University
MOLLENKOPF, William G., Educational Testing Service
MORRISON, Thomas F., Milton Academy
MUIRHEAD, Peter P., New York State Education Department
MYERS, Robert C., Educational Testing Service

ORLEANS, Beatrice S., Educational Testing Service
ORLEANS, Jacob S., The City College of New York
ORLEANS, Joseph B., George Washington High School
ORSHANSKY, Bernice, Educational Testing Service
OTTO, Eva M., American Occupational Therapy Association

PEASE, Katharine, Barnard College, Columbia University
PEDERSEN, Ruth A., Richardson, Bellows, Henry and Company
PERRY, William D., University of North Carolina
PETERSON, Donald A., Life Insurance Agency Management Association
PETERSON, Shailer A., American Dental Association
PLUMLEE, Lynnette, Educational Testing Service
POTTS, Edith M., The Psychological Corporation
PRITCHARD, Miriam C., Pratt Institute

REITZ, William, Wayne University
REYNOLDS, Mary E., Nichols Junior High School
RHODES, Catharine I., New York Public Schools
RICHARDSON, Marion W., Richardson, Bellows, Henry and Company
RINSLAND, Henry D., University of Oklahoma
RIVLIN, Harry N., Queens College
ROCK, Robert T., Fordham University
RODRIGUEZ-BOU, Ismael, University of Puerto Rico
ROHRER, John H., Yale University
ROHRS, John R., Staten Island, New York
RULON, Phillip J., Harvard University

SANDGREN, Dorothy C., Educational Testing Service
SCATES, Douglas E., American Council on Education
SCHAEFER, Willis C., University of Maryland
SCHRADER, William B., Educational Testing Service
SCHROEDEL, Edward C., International Business Machines Corporation
SCHULTZ, Doughas G., Educational Testing Service
SEASHORE, Harold G., The Psychological Corporation
SEGEL, David, U. S. Office of Education, Washington
SELOVER, Robert B., Prudential Insurance Company of America
SHARP, Catherine G., Educational Testing Service
SHAYCROFT, Marion F., National League of Nursing Education
SMITH, Allan B., University of Connecticut
SMITH, Denzel D., University of Maryland
SPANEY, Emma, Queens College
SPAULDING, Geraldine, American Institute for Research
SPENCER, Lyle M., Science Research Association
SPROULL, Lois, Standard Oil Company of New Jersey
STALNAKER, John M., Stanford University
STEPHAN, Frederick F., Princeton University
STEWART, Naomi, Educational Testing Service
STROMBERG, Eleroy L., Cleveland College
STUTZ, Frederick H., Cornell University
SULLIVAN, Richard H., Educational Testing Service
SYMONDS, Percival M., Teachers College, Columbia University

THIBALT, Paula, Educational Testing Service
THORNDIKE, Robert L., Teachers College, Columbia University
TRAVERS, Robert M. W., University of Michigan
TRAXLER, Arthur E., Educational Records Bureau
TRIGGS, Frances O., Committee on Diagnostic Reading Tests, Inc.
TUCKER, Ledyard R, Educational Testing Service
TUKEY, John W., Princeton University

TURNBULL, William W., Educational Testing Service

UPSHALL, Charles C., Eastman Kodak Company

VAUGHN, K. W., Harvard University and Educational Research Corporation

WALKER, Helen M., Teachers College, Columbia University

WANTMAN, Morey J., University of Rochester

WATT, R. R. G., University of Southern California

WESMAN, Alexander G., The Psychological Corporation

WHITNEY, Alfred G., Department of State

WOLMAN, Benjamin, Tel-Aviv, Israel

WOOD, Ben D., Columbia University

WOOD, Ray G., State Department of Education, Ohio

WRIGHTSTONE, J. Wayne, Board of Education, New York

ZUBIN, Joseph, Columbia University